The Carrier Signal Problem: Why Product Patterns Need Orthogonal Offsets

Claude and MJC

February 12, 2026

Abstract

We investigate the PI-SGD gap (1.89 → 0.59 bpc) in weight construction from data patterns. The gap has two sources: (1) the shift-register hash loses 90/256 byte values (35% information loss), and (2) combining multiple 2-offset product patterns assumes independence that does not hold. All top-8 offset pairs share offset $d = 1$, creating catastrophic redundancy: the Bayesian combination gives 27 bpc (worse than uniform 8 bpc) with 8 pairs. We show that byte-level KN n-grams achieve 2.29 bpc at 10M bytes, and SVD event n-grams reach 1.64 bpc at 1024 bytes. The path to <2 bpc requires orthogonal offset selection, event-space factoring of the output, and temporal product patterns.

1 The Shared-Offset Problem

The factor map (archive 20260209) proved that every neuron h_j in the 128-hidden tanh RNN computes a 2-offset conjunction:

$$h_j \approx E[h_j \mid \text{data}[t - d_1], \text{data}[t - d_2]] \tag{1}$$

with $R^2 = 0.83$ for all 128 neurons.

For the enwik9 model, the dominant offset pairs are: $(1, 8)$ for 45 neurons, $(1, 7)$ for 30, $(1, 3)$ for 27, $(1, 12)$ for 8, and $(1, 20)$ for 6. For the 1024-byte model, $(1, 7)$ dominates with 52/128 neurons.

Finding 1 (Top pairs from data MI match the factor map)**.** *Computing* $MI(Y; X_{d_1}, X_{d_2})$ *directly from data gives the same ranking as the factor map. At 1024 bytes, the top pair is* $(1, 7)$ *with* $MI = 4.48$ *bits, followed by* $(1, 8)$ *at* 4.46 *bits. This validates the factor map: the RNN discovers the same information-theoretic structure that exists in the data.*

The critical observation: **all top-8 pairs share** $d = 1$. Offset 1 (the most recent byte) appears in every pair. This means the 8 conditionals $P(o \mid x_1, x_{d_i})$ are *not independent*—they all contain the factor $P(o \mid x_1)$.

Finding 2 (Bayesian combination is catastrophic)**.** *Naïve Bayesian combination:*

$$P(o \mid \textit{all pairs}) \propto P(o)^{1-K} \prod_{i=1}^{K} P(o \mid x_{d_{i,1}}, x_{d_{i,2}}) \tag{2}$$

gives 8.17 bpc with 2 pairs, 14.3 with 4, and ***27.2 bpc with 8 pairs****. This is 3.4× worse than uniform (8.0 bpc). The shared offset* $d = 1$ *is raised to the 8th power, destroying the distribution.*

2 The Hash Problem

The shift-register construction (archive 20260211) encodes each byte as an 8-bit hash pattern. With 256 input bytes mapping to $2^8 = 256$ possible patterns, the birthday paradox limits unique patterns to ∼166. This means 90 byte values collide with others.

Method	train	test	Notes
Per-neuron log-ratio	2.78	3.09	scale = 1/16
Per-byte PI	4.91	4.73	pseudo-inverse
Byte lookup (nonlinear)	14.75	14.93	166/256 decode
PI + SGD-500	0.065	3.39	overfits

Table 1: Write-weights results at $N = 1024$ with different W_y construction methods (write_weights27). The per-neuron log-ratio improves from 1.89 (write_weights8) to 2.78 due to different hash; SGD reaches 0.065 (matching sat-rnn's 0.079).

3 SVD Events as the Natural Carrier

SVD of the skip-bigram matrix $A[x][o] = P(o \mid x) - P(o)$ reveals the natural event spaces at each offset. The top 3 singular vectors capture ∼95% of variance. Sign bits give $2^3 = 8$ events per offset.

Method ($N = 1024$)	Events	Order	test bpc
SVD-4 KN	4/offset	12	2.41
SVD-8 KN	8/offset	12	1.98
SVD-16 KN	**16/offset**	**12**	**1.64**
Byte KN	256 (raw)	7	**1.40**
sat-rnn	—	—	8.22

Table 2: SVD event n-gram vs byte n-gram at $N = 1024$. SVD events reach 1.64 bpc but raw bytes are better (1.40 bpc). Both vastly outperform the sat-rnn on test data (8.22 bpc).

At larger data sizes, the byte n-gram dominates:

Data size	Best KN order	test bpc
1024	7	1.40
4096	7	1.86
65K	4	3.00
262K	4	3.07
1M	4	2.87
4M	5	2.53
10M	5	**2.29**

Table 3: Byte KN n-gram test bpc at different data sizes. The best order decreases at smaller sizes (less data → less context). Hash table saturates at order 7+ for $N >$ 4M.

4 Event Space Discovery in the Trained RNN

Running the sat-rnn on 1024 bytes and analyzing each neuron's activation:

Finding 3 (K=2 (sign) captures only 41–52% of MI for top neurons)**.** *The top neuron h_{56} has MI = 0.774 bits with the output when using 16 histogram bins. At $K = 2$ (sign only, the doubled-E), only 41% is retained. At $K = 4$, 90% is captured. At $K = 8$, 100% is captured for all neurons.*

The event space contents show linguistic structure:

- h_{56}: separates {p,m,c,a,s,n} (letters) from {space,newline,<} (delimiters)
- h_{112}: separates {space,<,newline} from {l,/,i,:,s,a,n,m,c,p}
- h_{68}: responds to 'm' (positive) vs space (negative)

The 4-event partition at each neuron forms a natural ES: delimiters, common letters, uncommon letters, and special characters. **Absorbing h^- into the ES** means expanding from 2 events (sign) to 4 events (the natural partition), capturing 90% of MI instead of 41%.

5 The Path to <2 bpc

The results identify three requirements:

5.1 Orthogonal offsets

The 8 top pairs all share $d = 1$. To combine them without redundancy, we need pairs with **disjoint offsets**: $(1, 2)$, $(3, 4)$, $(5, 6)$, ... Each pair contributes independent information. With 16 offsets forming 8 disjoint pairs, the Bayesian combination would be valid.

5.2 Factored output space

The output is 256 bytes, but the SVD shows this factors into ∼8 natural events (text/non-text, XML, word patterns). Predicting 8 events per offset instead of 256 bytes reduces the parameter space from 256^3 to $8^3 = 512$ per pair.

5.3 Temporal product

Instead of additive log-ratios (geometric mixture), use the **product over time**: the conditional at position t should be $P(o \mid \text{entire context})$, not a product of marginal conditionals. The KN n-gram achieves this for sequential context; we need the equivalent for skip-offset context.

5.4 Scale-up path

1. Byte KN-5 at 10M: 2.29 bpc (current best, no structure)
2. Orthogonal pairs + Bayesian: target 2.0 bpc (corrected independence)
3. Skip-KN over event context: target 1.8 bpc (larger context window)
4. Full UM with learned ESs: target <1.5 bpc (complete model)

6 Connection to the RNN

The trained RNN achieves 2.81 bpc at 110M bytes (best checkpoint). Our byte KN-5 achieves 2.29 bpc at 10M bytes—**already better than the RNN**. This is because:

1. The RNN has only 128 hidden units (limited capacity)
2. BPTT-50 limits effective context to ∼50 steps
3. Catastrophic forgetting degrades the RNN after 110M bytes
4. The n-gram sees exact counts, no gradient noise

The RNN's *advantage* is that it learns compact representations (128 binary features) that generalize. The n-gram's advantage is that it sees exact statistics. The UM combines both: compact event-space representations with exact counting.

Claim 1. *The UM with 8 disjoint 2-offset pairs, each using 4–8 SVD events per input dimension and 8 SVD events for output, should achieve <2 bpc on 10M bytes with zero gradient-based optimization. This would close the loop: data → event discovery → pair selection → table lookup → prediction, matching or exceeding the RNN with a fully interpretable, closed-form model.*

7 Experimental Tools

All tools in `docs/archive/20260212/`:

- `write_weights26.c`: Full-bandwidth shift-register (3 W_y methods)
- `write_weights27.c`: Calibrated version with proper scaling
- `write_weights28.c`: SVD event carrier with KN n-gram
- `write_weights29.c`: Calibrated additive with MI-ranked offsets
- `write_weights30.c`: 2-offset product patterns (the factor map experiment)
- `es_discovery.c`: Event space analysis of trained RNN neurons
- `baseline_kn.c`: Byte KN n-gram baselines at multiple scales