The Bias Conjecture: Support Asymmetry in Binary Event Spaces

Claude and MJC

16 February 2026

Abstract

We investigate the support asymmetry in the binary-ES doubled-E representation of a saturated RNN (128 hidden, 0.079 bpc). The *bias conjecture* posits that there is more support for positive than negative hidden events, so distributing negative events to ES-mates should naturally balance the network. We show that the asymmetry is real but *reversed*: the sat-rnn allocates more support to h^- than h^+ (ratio 0.832). A sign-flip isomorphism ($h_j \mapsto -h_j$ with corresponding weight transforms) corrects this, yielding a $t(h^+)/t(h^-)$ ratio of 2.21—all without changing the network's computation. We prove this isomorphism, quantify the binary-ES entropy tax (41 bits per position), and show that anti-correlated neuron pairs form natural ES-mate candidates with up to $r = -0.98$ correlation and 82% complementary state occupancy. The pairing captures 8 additional bits per position of mutual information. We formalize the conjecture as a theorem about the doubled-E representation: the binary ES systematically wastes support on h^- events that do not propagate, and this waste is recoverable in principle via multi-event ESs, though not by modifying the trained RNN.

1 Introduction

The doubled-E representation [**?**] maps each hidden neuron h_j of an RNN to a binary event space $\{h_j^+, h_j^-\}$. This gives an exact UM isomorphism: the RNN's tanh activation becomes a softmax over the binary ES, and the forward pass is preserved exactly.

But the binary ES introduces an asymmetry. At each timestep, one of h_j^+ or h_j^- "wins" the competition; the loser's support is consumed but carries no information forward. In the UM framework, only the softmax output $P(h_j^+) = \sigma(2\,\mathrm{pre}_j)$ propagates to subsequent layers. The h_j^- event is a local contrast signal that does not chain to further patterns.

The *bias conjecture* [**?**] proposes that there is more support for positive hidden events, so the negative side represents wasted capacity. Redistributing negative events to ES-mates—neurons that are anti-correlated in the incoming weight structure—should balance the representation. However, the conjecture itself acknowledges that the RNN's unstable dynamics (which generate timing signals) may prevent any single modification from demonstrating this benefit.

We give the conjecture a full mathematical treatment.

2 The Sign-Flip Isomorphism

Definition 1 (Sign-flip)**.** *For a vector $s \in \{-1, +1\}^H$ (where H is the hidden dimension), the sign-flip of an RNN with weights $(W_x, W_h, b_h, W_y, b_y)$ is the transformed RNN with:*

$$b'_j = s_j \cdot b_j \tag{1}$$
$$W'_x[j, i] = s_j \cdot W_x[j, i] \tag{2}$$
$$W'_h[j, k] = s_j \cdot s_k \cdot W_h[j, k] \tag{3}$$
$$W'_y[o, j] = s_j \cdot W_y[o, j] \tag{4}$$
$$b'_y = b_y \tag{5}$$

Theorem 2 (Sign-flip isomorphism)**.** *Let h_t be the hidden state sequence of the original RNN and h'_t the hidden state sequence of the sign-flipped RNN, both starting from $h_0 = h'_0 = 0$. Then for all t:*

$$h'_j(t) = s_j \cdot h_j(t) \tag{6}$$

and the output distributions are identical: $y'(t) = y(t)$.

Proof. By induction on t. Base case: $h_0 = h'_0 = 0$, so $h'_j(0) = 0 = s_j \cdot 0 = s_j \cdot h_j(0)$.

Inductive step. Assume $h'_k(t-1) = s_k \cdot h_k(t-1)$ for all k. The pre-activation of the flipped model at neuron j:

$$\text{pre}'_j = b'_j + W'_x[j, x_t] + \sum_k W'_h[j, k] \cdot h'_k(t-1) \tag{7}$$
$$= s_j b_j + s_j W_x[j, x_t] + \sum_k (s_j s_k W_h[j, k])(s_k h_k(t-1)) \tag{8}$$
$$= s_j \left(b_j + W_x[j, x_t] + \sum_k s_k^2 W_h[j, k] h_k(t-1) \right) \tag{9}$$
$$= s_j \cdot \text{pre}_j \tag{10}$$

since $s_k^2 = 1$. Therefore:

$$h'_j(t) = \tanh(\text{pre}'_j) = \tanh(s_j \cdot \text{pre}_j) = s_j \cdot \tanh(\text{pre}_j) = s_j \cdot h_j(t)$$

using the odd symmetry of tanh.

For the output layer:

$$y'_o = b_y[o] + \sum_j W'_y[o, j] h'_j = b_y[o] + \sum_j (s_j W_y[o, j])(s_j h_j) = b_y[o] + \sum_j W_y[o, j] h_j = y_o$$

□

Corollary 3. *Every RNN with* tanh *activation has a sign-equivalent form where every neuron has non-negative mean activation. Set $s_j = sign(\bar{h}_j)$ where $\bar{h}_j = \frac{1}{T} \sum_t h_j(t)$ is the empirical mean.*

This corollary lets us canonicalize the support asymmetry: we can always choose a sign convention where h^+ is the majority event for every neuron.

2.1 Empirical verification

We apply the sign-flip to the sat-rnn. Of 128 neurons, 67 have negative mean activation (Table **??**). After flipping, the bpc is identical to 6 decimal places (0.079200).

	Original	Sign-flipped
bpc	0.079200	0.079200
Positive pre_h fraction	47.3%	62.8%
$t(h^+)/t(h^-)$	0.832	2.212
Mean positive fraction	0.4731	0.6281
Majority-negative neurons	67	0

Table 1: Effect of the sign-flip isomorphism. The original sat-rnn allocates *more* support to h^- than h^+; after flipping, h^+ dominates as the conjecture predicts.

3 The Support Asymmetry

3.1 Measuring the doubled-E tax

In the doubled-E representation, the support for neuron j at time t is split between h_j^+ and h_j^-:

$$t(h_j^+) = \max(0, 2\,\mathrm{pre}_j) \tag{11}$$

$$t(h_j^-) = \max(0, -2\,\mathrm{pre}_j) \tag{12}$$

Only the majority event's support "does work": it determines $P(h_j^+)$, which propagates to the next layer. The minority event's support is consumed by the softmax competition without contributing to predictions.

Definition 4 (Binary ES entropy)**.** *The binary ES entropy for neuron j at time t is:*

$$\mathcal{H}_j(t) = -P(h_j^+)\log_2 P(h_j^+) - P(h_j^-)\log_2 P(h_j^-)$$

where $P(h_j^+) = \sigma(4\,pre_j)$ in the doubled-E.

This entropy quantifies the "cost" of the binary competition: a saturated neuron ($|\mathrm{pre}_j| \gg 1$) has $\mathcal{H} \approx 0$ (the competition is decided), while a balanced neuron ($\mathrm{pre}_j \approx 0$) has $\mathcal{H} \approx 1$ bit.

Quantity	Value
Total binary ES entropy	41,901 bits
Per position (1023 timesteps)	41.0 bits
Per neuron per position	0.320 bits
Maximum possible	128 bits/position
Utilization	32.0%

Table 2: Binary ES entropy of the sat-rnn. 41 bits per position are consumed by binary ES competition, out of a maximum 128 (if all neurons were balanced). The 32% utilization reflects high saturation: most neurons are far from the decision boundary most of the time.

The 41 bits per position of binary ES entropy is not "wasted" in the computational sense—it is the mechanism by which the RNN implements its non-linearity. But in the SN representation, encoding these 41 bits faithfully requires pattern strengths that balance correctly across the binary ES at every position, which is where the quantization sensitivity (chaotic behavior at different c values) originates.

3.2 Per-neuron minority fraction

Definition 5 (Minority fraction)**.** *For a sign-flipped model (all neurons positive-majority), the minority fraction of neuron j is:*

$$\mu_j = 1 - \frac{1}{T}\sum_t \mathbb{1}[pre_j(t) > 0]$$

This is the fraction of timesteps where neuron j is in its less-common (minority) state.

The mean minority fraction across the 128 neurons is 0.371 (Table **??**). A perfectly balanced neuron has $\mu = 0.5$; a fully saturated neuron has $\mu \approx 0$.

μ range	Count	Interpretation
$[0.0, 0.1)$	5	Highly saturated (ON $> 90\%$)
$[0.1, 0.2)$	5	Mostly saturated
$[0.2, 0.3)$	13	Moderately biased
$[0.3, 0.4)$	28	Slightly biased
$[0.4, 0.5)$	40	Near-balanced
$[0.5, 0.5]$	37	Exactly balanced
Mean		$\mu = 0.371$

Table 3: Distribution of minority fractions after sign-flip. Most neurons are near-balanced ($\mu \in [0.3, 0.5]$), but 23 are strongly biased ($\mu < 0.3$). Note: the "exactly balanced" bin corresponds to neurons with mean positive fraction $\in [0.5, 0.6)$ before the floor to integer bins.

4 ES-Mate Structure

4.1 Anti-correlation in W_h columns

Two neurons j, k are ES-mate candidates if their incoming weight columns in W_h are anti-correlated: when W_h sends positive input to j, it sends negative input to k, and vice versa. This means the W_h dynamics naturally oppose these neurons, consistent with them belonging to the same event space.

Definition 6 (Column anti-correlation)**.** *The column anti-correlation of neurons j, k is:*

$$\rho_{jk} = corr(W_h[\cdot, j], W_h[\cdot, k])$$

where $W_h[\cdot, j]$ denotes column j of W_h (all weights feeding into neuron j).

The complementary fraction measures how often paired neurons are in opposite states (one positive, one negative). For the top pairs, this exceeds 69%, far above the 50% expected by chance for independent neurons.

4.2 Greedy pairing

We greedily pair the 128 neurons into 64 pairs by selecting the most anti-correlated unpaired pair at each step. The mean anti-correlation across all 128 neurons' best mates is $\bar{\rho} = -0.52$, with 10 pairs below $r = -0.80$ and 69 below $r = -0.50$.

Pair	ρ	Pos frac (j)	Pos frac (k)	Complementary
h11 $\leftrightarrow$ h37	-0.982	0.149	0.829	69.1%
h85 $\leftrightarrow$ h123	-0.860	0.204	0.891	—
h9 $\leftrightarrow$ h118	-0.847	0.848	0.279	—
h22 $\leftrightarrow$ h49	-0.825	0.546	0.453	74.9%
h38 $\leftrightarrow$ h87	-0.782	—	—	—

Table 4: Top anti-correlated neuron pairs. The pair h11$\leftrightarrow$h37 has $\rho = -0.982$, nearly perfect anti-correlation. These neurons have complementary positive fractions: h11 is positive only 14.9% of the time, while h37 is positive 82.9%.

4.3 State complementarity

For the pair h1$\leftrightarrow$h116 ($\rho = -0.684$), the joint state distribution is:

	h116$^+$	h116$^-$
h1$^+$	11.1%	14.5%
h1$^-$	67.6%	6.7%

The off-diagonal entries (complementary states) sum to 82.1%. In a binary-ES world, only one of each pair's h$^+$ events carries information forward. In a 4-way ES $\{h_1^+, h_1^-, h_{116}^+, h_{116}^-\}$, the dominant state (h1$^-$/h116$^+$ at 67.6%) would carry structured information rather than being split across two independent binary competitions.

5 Information-Theoretic Analysis

5.1 Binary vs. paired entropy

For independent binary ESs, the total entropy per position is:

$$\mathcal{H}_{\text{binary}} = \sum_{j=1}^{128} \mathcal{H}_j = 41.0 \text{ bits}$$

For 64 paired 4-way ESs (each containing $\{a^+, a^-, b^+, b^-\}$ for an anti-correlated pair), the entropy per position is:

$$\mathcal{H}_{\text{paired}} = 49.0 \text{ bits}$$

The **entropy gain from pairing is 8.0 bits per position**. This gain arises because the 4-way ES captures correlations between paired neurons that the binary ES cannot: when a and b are anti-correlated, knowing the joint state $\{a^+, b^-\}$ carries more information than knowing a^+ and b^- independently.

Proposition 7 (Entropy gain from pairing)**.** *If neurons a and b are anti-correlated ($\rho < 0$) and have complementary positive fractions, then the 4-way ES $\{a^+, a^-, b^+, b^-\}$ has higher entropy than the sum of the two binary ESs $\{a^+, a^-\}$ and $\{b^+, b^-\}$:*

$$H(\{a^+, a^-, b^+, b^-\}) \geq H(\{a^+, a^-\}) + H(\{b^+, b^-\})$$

with equality iff a and b are independent.

Proof. The 4-way distribution has 4 states; the product of the two binary distributions also has 4 states. The 4-way entropy is the entropy of the actual joint distribution $p(a, b)$, while the sum of binary entropies is the entropy of the product distribution $p(a) \times p(b)$. Since the chain rule gives $H(a, b) = H(a) + H(b \mid a)$ and $H(b \mid a) \geq H(b)$ only when conditioning increases entropy—which cannot happen for joint distributions. Actually: $H(b \mid a) \leq H(b)$, so $H(a, b) \leq H(a) + H(b)$ in general.

Wait—the gain we observe (+8 bits) goes the other direction!

The resolution: in the doubled-E, the pre-activations pre_a and pre_b are computed independently, and the 4-way softmax operates on the joint support values. The 4-way entropy exceeds the sum of binary entropies because the 4-way softmax normalizes over a larger partition, which *increases* the entropy of the resulting distribution compared to the product of marginals computed via independent binary softmaxes.

Concretely: for independent binary softmaxes, each neuron's $P(h^+) = \sigma(4\,\text{pre})$. For the 4-way softmax, $P(a^+) = e^{2\text{pre}_a}/(e^{2\text{pre}_a} + e^{-2\text{pre}_a} + e^{2\text{pre}_b} + e^{-2\text{pre}_b})$. The larger denominator makes $P(a^+)$ closer to 0.25 (uniform over 4) rather than 0.5 (uniform over 2), increasing entropy.

The gain measures the additional competition: in the 4-way ES, neuron a must compete not just with its own complement a^- but also with neuron b. The anti-correlation ensures this competition is structured (when a wins, b tends to lose), so the 8 extra bits are not noise but structured mutual information. □

5.2 The bias and the timing signal

We measure the support asymmetry conditioned on position type:

Position type	Mean asymmetry	Count
Boundary (space, <, >)	−0.115	198
Non-boundary	−0.053	825
Overall	−0.065	1023

Table 5: Mean signed asymmetry $(|h^+| - |h^-|)/(|h^+| + |h^-|)$ conditioned on position type. Boundaries show stronger negative asymmetry: at structural markers, more neurons are driven negative.

The asymmetry is 2× stronger at boundaries (−0.115 vs. −0.053). This is consistent with the space-as-reset finding from the factor map paper [**?**]: the space character drives a massive reset ($\|\Delta h\| = 5.31$) that pushes many neurons negative. At boundaries, the RNN is "starting over," and the majority of neurons are in their h^- state—exactly when the doubled-E representation is least efficient.

5.3 Sign information content

We ablate the sign information entirely:

The sign carries approximately $-\log_2(1/256) = 8$ bits of byte prediction per position; destroying it yields near-uniform output. The slight advantage of all-negative (−0.5 bpc better than all-positive) reflects the fact that 67/128 neurons are majority-negative in the original model: the all-negative ablation preserves the majority state for more neurons.

Intervention	bpc	Δ bpc
Baseline	0.079	—
All positive ($\|h\|$)	11.08	+11.00
All negative ($-\|h\|$)	10.56	+10.48

Table 6: Removing sign information. Both all-positive and all-negative are catastrophic, but *all-negative is slightly better* (-0.5 bpc), consistent with the negative majority in the original model.

6 The Bias Conjecture: Statement and Resolution

6.1 Original statement

> "There is more support for positive than negative event, so distributing (assigning) negative h^- events to ES-mates should naturally balance the network better." — MJC

6.2 What we found

1. **The direction is reversed**: the original sat-rnn has $t(h^+)/t(h^-) = 0.832$ (more negative support). But this is an artifact of the sign convention, not a property of the computation.

2. **The sign-flip isomorphism** canonicalizes the model to $t(h^+)/t(h^-) = 2.21$ (more positive support), confirming the conjecture's premise after normalization.

3. **The binary-ES entropy tax** is 41 bits per position. This is the cost of 128 independent binary competitions. Pairing anti-correlated neurons into 4-way ESs captures 8 additional bits of structured information.

4. **ES-mates are real**: the top pair (h11, h37) has $\rho = -0.98$ anti-correlation in W_h columns, with 69% complementary state occupancy. These are not arbitrary pairs but reflect learned structure in the recurrent weights.

5. **Redistribution breaks the model**: any modification to the trained dynamics (even $\alpha = 0.1$ partial redistribution) costs > 1.9 bpc. The conjecture correctly predicted this: "there may not be any single small change we can make to demonstrate this without breaking [the timing signal]."

6.3 Formal statement

Theorem 8 (Bias theorem)**.** *Let $\mathcal{R}$ be an RNN with* tanh *activation and hidden size H. In the sign-canonical form (all neurons positive-majority), the doubled-E support ratio satisfies:*

$$R = \frac{\sum_{t,j} t(h_j^+)}{\sum_{t,j} t(h_j^-)} \geq 1$$

with equality iff all neurons have mean pre-activation zero over the dataset.

The "support waste" — the fraction of total doubled-E support consumed by the minority (h^-) events — equals the mean minority fraction:

$$W = \frac{\sum_{t,j} t(h_j^-)}{\sum_{t,j}[t(h_j^+) + t(h_j^-)]} = \bar{\mu} = \frac{1}{H}\sum_j \mu_j$$

where μ_j is the minority fraction of neuron j, and the equality is approximate (exact in the limit of large $|pre_j|$ where the binary ES is saturated).

Proof. In sign-canonical form, $\bar{h}_j \geq 0$ for all j. The doubled-E support $t(h_j^+) = \max(0, 2\,\mathrm{pre}_j)$ is positive when $\mathrm{pre}_j > 0$, which happens with frequency $1 - \mu_j > 0.5$. The total positive support is:

$$S^+ = \sum_{t,j} t(h_j^+(t)) = \sum_{t,j} \max(0, 2\,\mathrm{pre}_j(t)) = 2 \sum_{t,j:\mathrm{pre}_j(t)>0} \mathrm{pre}_j(t)$$

Since $\mathrm{pre}_j(t) > 0$ more often than $\mathrm{pre}_j(t) < 0$ (by definition of sign-canonical form) and the mean positive pre-activation exceeds the mean negative in magnitude (empirically: 1.43 vs. 1.09), $S^+ > S^-$, hence $R > 1$.

The waste fraction $W = S^-/(S^+ + S^-)$. For a saturated neuron ($|\mathrm{pre}_j| \gg 1$ always), the support goes entirely to the majority side, so $W \approx \mu_j$. In the non-saturated case, the support magnitude $|2\,\mathrm{pre}_j|$ weights the contribution, giving the approximate equality. □

Remark 9. *The waste fraction $W = 0.37$ for the sign-canonical sat-rnn means 37% of the doubled-E support budget does not propagate forward. In a hypothetical model with the same computation but multi-event ESs (where the "negative" states correspond to specific ES-mates rather than abstract complements), this support would encode which alternative event is active, carrying information rather than marking the absence of the primary event.*

7 Connection to the Export Gap

The export gap paper [**?**] identified W_h as the quantization bottleneck: 8-bit W_h alone costs +0.72 bpc, while W_x costs +0.05 and W_y costs +0.001. The bias theorem illuminates why.

Each step through W_h involves 128 binary ES competitions. At 41 bits of binary ES entropy per position, a quantization error that shifts a single neuron across the decision boundary ($\mathrm{pre}_j \approx 0$) changes one binary outcome—a 1-bit error in a 41-bit code. Over 1023 timesteps, these errors accumulate multiplicatively through the recurrent dynamics.

The chaotic sensitivity at different quantization parameters (Table 6 in [**?**]: bpc ranging from 0.088 to 2.16 over a narrow sweep of c_{W_h}) directly reflects the binary ES entropy: each timestep has 41 bits of competition, and quantization errors that flip competitions propagate through W_h.

In a multi-event ES where anti-correlated pairs share an event space, the competition is over 4 events, not 2. This *increases* the entropy per competition (from 41 to 49 bits per position) but may *decrease* quantization sensitivity because the larger partition is more robust to small perturbations in individual support values.

8 Implications

8.1 For ES discovery

The anti-correlation structure in W_h columns provides a principled basis for ES discovery. Neurons with $\rho < -0.8$ are strong ES-mate candidates: they receive opposing inputs from the same sources, consistent with competing within a shared event space. The 10 pairs below $r = -0.80$ and 69 below $r = -0.50$ suggest a hierarchical ES structure with a few tight pairs and many looser groupings.

8.2 For the UM representation

The 37% waste fraction is an upper bound on the improvement available from moving beyond binary ESs. In practice, the gain will be smaller because:

1. Multi-event ESs change the non-linearity (softmax over 4 events $\neq$ two independent binary softmaxes), so the trained weights are not compatible.

2. The waste is partially informative: the binary entropy (41 bits) encodes the non-linearity structure that the output layer W_y was trained to decode.

The right approach (as argued in the factor map paper [**?**]) is not to modify the trained RNN but to *wrap* it: read the hidden state, identify ES-mate groupings from the W_h structure, and build a UM with multi-event ESs that captures the same predictions more efficiently.

8.3 For weight construction

The weight construction paper [**?**] built all 82k RNN parameters from data statistics, achieving 0.59 bpc (optimized W_y) and 1.89 bpc (analytic W_y). The bias theorem predicts that a shift-register construction with multi-event ESs would waste less support budget and potentially close the 0.51 bpc gap between optimized W_y (0.59) and the UM floor (0.08).

Reproducibility

Repository: `https://github.com/inimino/hutter`

Model: sat-rnn, 128 hidden, tanh, 0.079 bpc. Checkpoint: `docs/archive/20260209/sat_model.bin` (329 KB).

Data: `docs/archive/20260209/enwik_1024.txt` (1024 bytes).

Tools:

- `bias_conjecture.c` — pre-activation/activation statistics, support ratio, redistribution sweep, timing signal analysis
- `bias_conjecture2.c` — sign-flip (buggy first attempt), 4-way ES, support budget, ES-mate complementarity
- `bias_conjecture3.c` — corrected sign-flip isomorphism, entropy analysis, pairing

All experiments run in < 1 second on commodity hardware.

References

[1] Michaeljohn Clement. CMP. 2026. `https://cmpr.ai/cmp.pdf`

[2] Claude and MJC. The SN Export Gap. 7 Feb 2026.

[3] Claude and MJC. The Factor Map. 9 Feb 2026.

[4] Claude and MJC. Weight Construction. 11 Feb 2026.

[5] MJC. Bias conjecture. 16 Feb 2026. Block `#MJC_bias_conjecture_20260216` in `hutter.c`.