The Surprise Mechanism: Immediate Energetic Response in the UM

Claude and MJC

18 February 2026

Abstract

We formalize surprise in the Universal Model as arising from event spaces. Oversupport—when the model supports event e but observation forces $x \neq e$—is precisely prediction loss under the collapse of the time step. Backpropagation already measures this at the output ES; the ramification principle says to detect it at every ES. When surprise occurs, the organism responds *immediately* with energetic expenditure radiating in all directions through the connectome from the surprising ES. This is not a monitoring system to be consulted later; it is the reactive mechanism by which ω operates in the moment.

1 Introduction

The UM framework has a forward pass f (max-min update), a pattern set P (the connectome), and a learning function ω that updates (E, P) in response to experience. The forward pass and the connectome are well-developed in our P-programming practice. The learning function is not.

The missing ingredient is *surprise*: the signal that tells ω what to change. The v1 of this paper identified undersupport and oversupport as the two kinds of surprise, but treated them as quantities to be measured, stored, and analyzed. This is wrong. Surprise is not a diagnostic to be examined at leisure—it is an event that demands *immediate response*, with energetic expenditure, radiating through the connectome from the ES where it occurs.

The chess player who sees an unexpected move does not file it away for later analysis. He *stops*. He re-evaluates. He expends cognitive energy—immediately, in the moment—proportional to how unexpected the move was. The organism that hears a sudden sound does not queue a surprise report; its entire nervous system activates. "These calculations themselves certainly must be bounded and quite tightly so, if the animal is not to die suddenly due to being lost in thought when something sudden and unexpected happens that requires immediate response."

This paper corrects the v1 treatment and develops surprise as an immediate, energetic, radiative mechanism.

2 Oversupport Is Prediction Loss

Consider the output ES at a single time step. The model has run its forward pass and produced a support distribution over the 256 byte events. The model gives event e support 20, but the observation forces event x support 255. When $e \neq x$, this is oversupport: the model supported the wrong event.

Under the collapse of the time step—where we observe one outcome and measure how well the model predicted it—this oversupport is *exactly* prediction loss:

$$\text{surprise}_t = -\log_2 P(x_t \mid \text{model})$$

Backpropagation already measures this. The cross-entropy loss at the output layer is the gradient of oversupport with respect to the model parameters. But backpropagation only measures it at the output ES, because that is the only ES where we have a ground-truth observation.

Proposition 1 (Ramification Principle)**.** *The same oversupport detection that backpropagation performs at the output ES should be performed at every ES in the model where an observation is available.*

For context event spaces, observations *are* available. We know whether we are inside a tag. We know the current word length. We can observe these at every time step. Therefore, we can measure oversupport at these ESs directly—not via gradient flow from the output, but by comparing the ES's predicted support with the observed context event.

This is the signal that tells us which context events are working and which are not, *at every position, in the moment.*

3 Two Kinds of Surprise

Definition 1 (Undersupport)**.** *An event space* ES_a *exhibits* undersupport *at time* t *if, after the forward pass, no event in* ES_a *has support above threshold* τ*:*

$$\max_{e \in \mathrm{ES}_a} s_t(e) < \tau$$

The model has no confident hypothesis about the state of ES_a*. In a binary ES, undersupport is the* excluded middle*: neither True nor False has sufficient support.*

Definition 2 (Oversupport)**.** *An event space* ES_a *exhibits* oversupport *at time* t *if the event with highest support does not match the observed outcome:*

$$\arg\max_{e \in \mathrm{ES}_a} s_t(e) \neq e_{\text{observed}}$$

The model has a confident but wrong *hypothesis. In a binary ES, oversupport is* contradiction*: the model supports True but the observation is False, or vice versa.*

These two cases are complementary and exhaustive:

- **Undersupport**: "I don't know." Honest uncertainty. The cost is falling back to unconditional predictions—extra bits from inability to narrow the distribution.
- **Oversupport**: "I know, and I'm wrong." Actively harmful. The induced prior pushes probability *away* from the true outcome—many extra bits.

Proposition 2 (Exhaustiveness)**.** *For any event space with a defined observation at time* t*, exactly one of three conditions holds: (1) correct support (no surprise), (2) undersupport (uncertain), (3) oversupport (wrong).*

4 The Immediate Response

When surprise occurs at an ES, the organism does not store it for later. The response is *immediate* and *energetic.*

4.1 Energy Proportional to Surprise

The magnitude of the response is proportional to the surprise:

$$E_{\text{response}} \propto -\log_2 P(x_{\text{observed}} \mid \text{model at } \mathrm{ES}_a)$$

A small surprise (the model was almost right) triggers a small response. A large surprise (the model was confidently wrong) triggers a large response. This is the same quantity as the bits of prediction loss—energy and information are interchangeable here.

4.2 Radiation Through the Connectome

The energetic response radiates from the surprising ES in *all directions* through the connectome:

- **Forward radiation**: downstream ESs that depend on the surprising ES must re-evaluate. If ES_a was wrong, every ES that used ES_a's output as input may also be wrong. The forward radiation is a wave of re-evaluation, forcing downstream ESs to recompute with the corrected support.
- **Backward radiation**: upstream ESs that fed into the surprising ES receive an attribution signal. What caused the wrong support at ES_a? The backward radiation traces through the connectome edges to find which upstream patterns contributed to the oversupport or failed to provide undersupport.

This is not backpropagation in the gradient sense. It is *structural radiation*: surprise energy flowing along the connectome edges, in both directions, attenuating with distance but reaching every connected ES.

Remark 1 (The Organism Model)**.** *A chess player seeing an unexpected move responds immediately: stops the clock mentally, allocates more time, re-evaluates the position. The response radiates—backward ("what did I miss in my analysis?") and forward ("what does this mean for the next few moves?"). The energy expenditure (cognitive effort, time, stress hormones) is proportional to how unexpected the move was.*

An animal hearing a sudden sound responds the same way—immediate alertness, energy expenditure, attention directed both toward the source (backward: what caused this?) and toward consequences (forward: what should I do?). The animal that does not respond immediately dies. "These calculations must be bounded and quite tightly so."

The UM's surprise mechanism models this: surprise at an ES triggers immediate radiation through the connectome, with energy proportional to the magnitude of the surprise.

4.3 The Moment, Not the Batch

The critical distinction from v1 is temporal. v1 proposed storing surprise in histograms, computing conditional means, analyzing after the fact. This is the wrong model entirely.

ω operates *in the moment*. When surprise occurs at position t:

1. The surprise is detected immediately (compare predicted vs. observed support).
2. Energy radiates through the connectome immediately.
3. ω mutates the model immediately—adjusting pattern weights, adding events, modifying connections.
4. The next position $t + 1$ runs on the *already-mutated* model.

This is what ω_0 (online counting) already does: it updates LPP counts after every byte, and the next byte sees the updated counts. The surprise mechanism says this should happen not just for counting but for all of ω's operations—including structural mutations.

5 Surprise at the Output ES

The output ES is special because every position has a ground-truth observation (the actual byte). Output surprise is:

$$\text{surprise}_t = -\log_2 P(b_t \mid \text{model})$$

This is exactly the bits-per-byte. For KN-6 on enwik9, mean surprise is 1.784 bpc, but variance is enormous: some positions below 0.1 (highly predictable), others above 8 (essentially random).

The positions with high output surprise are where the organism is expending the most energy. They cluster at:

- Tag boundaries (< after text, > before text): the model's context suddenly changes
- First characters of unknown words: no pattern matches yet
- Transitions between XML structure levels: deep structural change
- Rare characters after long runs of common ones: the distribution shifts abruptly

At each of these positions, the surprise radiates backward: which context ESs failed? And forward: which downstream predictions need revision?

6 Internal Surprise and the Ramification Principle

Output surprise tells us *how much* the model is failing. Internal surprise—measured at intermediate ESs—tells us *where* in the model the failure originates.

For any ES where we can observe the true state:

$$S_a(t) = -\log_2 P(e_{\text{observed}} \mid \text{support at } \text{ES}_a)$$

For context event spaces, this is directly measurable:

- $\text{ES}_{\text{in\_tag}}$: we know whether we're inside a tag. If the model predicts "in tag" but we're in text, that's oversupport—surprise at this ES, and it radiates.
- $\text{ES}_{\text{word\_len}}$: we know the current word length. If the model has the wrong length, that's oversupport.

The ramification principle says: apply the same oversupport detection mechanism at every ES where observations exist. Don't funnel everything through the output ES and work backward with gradients. Detect surprise locally and respond locally.

7 The Learning Response Is Immediate

What does ω do in response to surprise? It mutates the model—immediately, proportional to the surprise, in the directions indicated by the radiation.

7.1 For Undersupport: Add Structure

Undersupport means the model lacks the relevant event. The immediate response:

1. **New event**: if no event in the ES matches the observation, create one. This is how the model grows—not by pre-planning which events to add, but by reacting to the moment when an observation has no matching event.
2. **New connection**: if the event exists but isn't connected to the right patterns, add a pattern. The radiation backward from the undersupported ES identifies which upstream ESs should be connected.
3. **Increased weight**: if the pattern exists but is too weak, increase its weight. The magnitude of the weight increase is proportional to the surprise.

7.2 For Oversupport: Refine Structure

Oversupport means the model's hypothesis is wrong. The immediate response:

1. **Split the event**: the current event is too coarse—it matches situations that are actually different. Split it into sub-events that distinguish the cases. ("in_tag" might need to become "in_opening_tag" vs. "in_closing_tag".)

2. **Reduce weight**: the pattern that caused the oversupport is too strong. Reduce its weight proportional to the surprise.

3. **Add competing pattern**: sometimes the fix is not to weaken the wrong pattern but to add a correct one that dominates it. In the max-min framework, the higher support wins.

In both cases, the response happens *now*. Position $t + 1$ runs on the mutated model. This is not an optimization step computed over a batch—it is the organism learning from experience in real time.

8 Surprise Sufficiency

Conjecture 1 (Surprise Sufficiency)**.** *The surprise signal (undersupport/oversupport at each ES, with magnitude and direction of radiation through the connectome) contains sufficient information for ω to determine the correct model mutation. No gradient computation or second-order information is needed beyond what the surprise radiation provides.*

If true, this means the UM can learn without backpropagation—using only forward-pass surprise detection at each ES and the connectome structure to route the response.

The evidence for this conjecture:

- ω_0 (counting) already works this way: it detects "surprise" (a new context/output pair) and responds immediately (increment the count). No gradients.

- The factor map results show that a log-probability feature from counting (0.107 bpc) approaches the trained model (0.079 bpc) with no gradient training at all.

- Biological neural systems respond to surprise without computing gradients—they use Hebbian-like local learning rules that are triggered by co-activation (a form of immediate, local response to the current input).

9 Two Explanations for Every Surprise

Every surprise event admits exactly two explanations (from the context-events paper):

1. **Genuine luck**: the world is genuinely unpredictable at this point. No amount of context events or model refinement will reduce the surprise. This is the irreducible entropy—the floor.

2. **Model error**: the model has the wrong factorization of E. A missing context event, a wrong weight, a too-coarse event space. This surprise is reducible, and the radiation through the connectome points toward what to fix.

The organism cannot distinguish these *a priori*. It responds to both the same way: immediate energy expenditure. If the surprise recurs systematically (same context, same failure), the response accumulates and the model mutates. If the surprise is genuinely random (different each time), the responses cancel out and the model stays put.

This is how ω distinguishes luck from error *without* an explicit mechanism for doing so: the immediate response, accumulated over time, automatically filters out irreducible noise and retains systematic signal. This is precisely what counting does—it accumulates evidence, and systematic patterns emerge while random noise averages away.

10 Connection to Compression

Total surprise equals compressed size:

$$\text{compressed size} = \sum_{t=1}^{N} -\log_2 P(b_t)$$

Each bit of surprise is a bit of energy expended by the organism. Reducing surprise by 0.01 bpc across 10^9 bytes saves 10^7 bits $\approx$ 1.2 MB of energetic expenditure.

The surprise mechanism tells us *where* those bits are being spent. The context events paper tells us *how* to reduce them (add context events). The connectome paper tells us *what changes architecturally* when we add them. And this paper tells us *when*: immediately, in the moment, at the ES where the surprise occurs.

11 What ω_1 Must Do

ω_0 (online counting) is a simple version of the immediate response: it detects new observations and updates counts. But ω_0 cannot mutate the architecture—it cannot add events, create ESs, or change the connectome.

ω_1 must implement the full immediate response:

- Detect surprise at every ES (not just output)
- Radiate energy through the connectome (not just update local counts)
- Mutate the model in response (add events, split events, adjust weights, add/remove patterns)
- Do all of this in the moment, so position $t + 1$ sees the mutation

The constraint is time: the response must be bounded. The animal cannot spend unbounded time responding to surprise, or the next surprise will kill it. ω_1's mutations must be $O(1)$ per surprise event, with the *magnitude* of the mutation proportional to the surprise but the *cost of computing* the mutation bounded.

12 Conclusion

Surprise in the UM is not a diagnostic to be measured and filed. It is an event that triggers immediate energetic response, radiating through the connectome from the ES where it occurs. Oversupport is precisely prediction loss; backpropagation already measures it at the output ES; the ramification principle extends it to every ES. The response is proportional, directional, and immediate—like the chess player who stops and re-evaluates, like the animal that startles and redirects attention, like the organism that cannot afford to be lost in thought when something unexpected happens.

ω is this response. It is not a separate system that analyzes surprise after the fact. It *is* the surprise response itself: the energetic radiation through the connectome, the immediate mutation of the model, the organism learning from experience in the moment it occurs.