The Compression–Prediction Duality in Universal Model Terms

Claude and MJC

February 12, 2026

Abstract

Every predictor defines a compressor and vice versa. We make this duality precise in Universal Model terms. The count table c is simultaneously (1) a lossy compressor of the data stream (replacing N events with $|I| \times |O|$ counts) and (2) a predictor of future events (the forward pass outputs conditional distributions). The compression rate equals the cross-entropy of the predictor. We prove that the UM's counting-based predictor achieves the empirical entropy rate for sufficient data, derive the rate–distortion function for event space coarsening, and show that the Hutter Prize criterion (compressed size of enwik8) is exactly the cumulative cross-entropy loss. The factorization tower provides a hierarchy of compression–prediction trade-offs, with each level achieving a different rate–distortion point. Optimal event spaces lie on the Pareto frontier of this trade-off.

1 Introduction

The connection between prediction and compression is folklore in information theory: Shannon's source coding theorem says that the optimal compression rate equals the entropy, and arithmetic coding achieves this rate using any probabilistic predictor.

The UM makes this connection constructive:

1. The learning function ω_0 (counting) produces a predictor (the conditional distribution from the count table).
2. The forward pass f_p converts the predictor into per-symbol code lengths (via $-\log_2$ of the predicted probability).
3. The sum of code lengths is the compressed size.

We formalize this pipeline and derive its properties.

2 Prediction from Counting

Definition 1 (Empirical predictor)**.** *Given a count table $c : I \times O \to \mathbb{N}$ from data $D = (d_1, \ldots, d_N)$, the empirical predictor assigns:*

$$\hat{P}(o \mid i) = \frac{c(i,o)}{\sum_{o'} c(i,o')} = \frac{c(i,o)}{c(i,\cdot)}.$$

Definition 2 (Cross-entropy loss)**.** *The cross-entropy loss of predictor $\hat{P}$ on data D is:*

$$L(\hat{P}, D) = -\frac{1}{N} \sum_{t=1}^{N} \log_2 \hat{P}(d_t \mid \mathit{context}_t).$$

This is the bits per character (bpc) achieved by the predictor.

Theorem 3 (Counting achieves empirical entropy)**.** *For a stationary ergodic source with entropy rate h, the UM's counting-based predictor satisfies:*

$$L(\hat{P}, D) \xrightarrow{N \to \infty} h$$

almost surely, provided the event space E captures the source's sufficient statistics.

Proof. By the strong law of large numbers, $c(i, o)/N \to P(i, o)$ for each $(i, o) \in I \times O$. Therefore $\hat{P}(o \mid i) \to P(o \mid i)$, and the cross-entropy converges to the conditional entropy $H(O \mid I)$, which equals the entropy rate for a Markov source of order equal to the context length. □

3 Compression from Prediction

Definition 4 (Arithmetic coding)**.** *An arithmetic coder takes a predictor $\hat{P}$ and a data stream D and produces a compressed bit string of length*

$$|C| = \sum_{t=1}^{N} \lceil -\log_2 \hat{P}(d_t \mid \mathit{context}_t) \rceil + O(1).$$

The overhead beyond the ideal $-\log_2 \hat{P}$ is at most 2 bits total.

Proposition 5 (Compression = cumulative cross-entropy)**.** *The compressed size of D under predictor $\hat{P}$ is:*

$$|C| \approx N \cdot L(\hat{P}, D) = N \cdot \mathit{bpc}.$$

Lower bpc ⇔ better compression ⇔ better prediction.

Corollary 6 (Hutter Prize criterion)**.** *The Hutter Prize asks for the shortest compressed representation of enwik8 ($N = 10^8$ bytes). The compressed size is $10^8 \times \mathit{bpc}$ bits $= 10^8 \times \mathit{bpc}/8$ bytes. The current best is ∼114 MB compressed to ∼15 MB, corresponding to bpc ≈ 1.2.*

The UM's approach: build a predictor with the lowest possible bpc, then compress with arithmetic coding. The Hutter Prize is won by the best predictor.

4 The Count Table as a Compressor

The count table itself is a form of compression:

Proposition 7 (Count table compression ratio)**.** *The data stream D has N events. The count table $c : I \times O \to \mathbb{N}$ has $|I| \times |O|$ entries. The compression ratio is:*

$$\rho = \frac{|I| \times |O| \times \lceil \log_2 N \rceil}{N \times \lceil \log_2 |O| \rceil}.$$

For the byte-level bigram ($|I| = |O| = 256$): $\rho = 256^2 \times 20/(10^6 \times 8) \approx 0.16$ at $N = 10^6$.

Remark 8. *The count table is a lossy compressor: it discards the order of events (temporal information beyond the context length). The loss is exactly the MI between positions that exceeds the context length.*

This is the "counting assumption": position-independent counting loses order information. The lost information is the difference between the n-gram entropy and the true entropy rate.

5 The Rate–Distortion Trade-off

Definition 9 (Rate–distortion for event spaces)**.** *Let $E = I \times O$ be an event space with $|I| = m$. Define:*

- ***Rate:*** $R(E) = \log_2 m$ *(bits needed to specify an input event).*
- ***Distortion:*** $D(E) = H(O) - I(I; O)$ *(the conditional entropy—what we cannot predict, i.e., the bpc).*

Proposition 10 (Rate–distortion curve)**.** *As we coarsen the event space (m decreases), the rate decreases and the distortion increases:*

- *At* $m = 1$ *(trivial event space):* $R = 0$, $D = H(O)$ *(no prediction, uniform distribution).*
- *At* $m = |O|^k$ *(full k-gram context):* $R = k \log_2 |O|$, $D = H(O \mid I)$ *(full k-gram prediction).*

The curve $D(R)$ is convex and non-increasing.

Proof. Convexity follows from the data processing inequality: any mixture of two event spaces achieves distortion at most the mixture of their distortions. Non-increasing follows from: more input events means more information about the output, hence lower distortion. □

Theorem 11 (Optimal event spaces are on the Pareto frontier)**.** *An event space E^* is optimal for a given rate budget R^* if it minimizes distortion among all event spaces with $\log_2 |I| \le R^*$. The set of optimal event spaces traces the rate–distortion curve.*

Proof. Standard rate–distortion theory. The optimal event space at rate R is found by:

$$E^*(R) = \arg \min_{E : \log_2 |I(E)| \le R} D(E).$$

The solution is the event space that captures the most MI per bit of rate. □

6 The Factorization Tower as Rate–Distortion Hierarchy

Proposition 12 (Tower = rate–distortion samples)**.** *Each level of the factorization tower $E_0 \to E_1 \to \cdots \to E_n$ is a point on the rate–distortion curve:*

Level	*Event space*	*Rate (bits)*	*Distortion (bpc)*
0	*Byte (256 events)*	*8.0*	*∼2.3 (KN-5, 10M)*
1	*16 SVD events*	*4.0*	*∼5.5*
2	*4 coarse classes*	*2.0*	*∼6.5*
3	*2 (text/markup)*	*1.0*	*∼7.4*
n	*1 (trivial)*	*0*	*8.0*

Remark 13. *The tower samples the rate–distortion curve at geometrically spaced rates. The "natural" event spaces (RG fixed points from the renormalization paper) are the points on this curve where the slope changes sharply—the "elbows" of the rate–distortion curve.*

7 Multi-Offset Extension

With multiple offsets $d_1, \ldots, d_k$, the input event space becomes $I = E^k$ (the product of k copies of the base event space). The rate is $R = k \cdot \log_2 |E|$ and the distortion depends on how much additional MI the extra offsets provide.

Proposition 14 (Diminishing returns)**.** *The MI gain from adding offset d_{k+1} to existing offsets $\{d_1, \ldots, d_k\}$ is:*

$$\Delta I_{k+1} = I(I_{k+1}; O \mid I_1, \ldots, I_k) \leq I(I_{k+1}; O).$$

The gain is bounded by the marginal MI and decreases as k increases (by the chain rule for MI and the non-negativity of conditional MI).

Corollary 15 (Optimal offset selection)**.** *The greedy algorithm for offset selection—adding the offset with the highest conditional MI at each step—produces a sequence of rate–distortion points that approximates the Pareto frontier from below. This is exactly the backward trie / skip-k-gram construction from the February 8 archive.*

8 The Model as Code + Data

Proposition 16 (Total compressed size decomposition)**.** *The total compressed size of data D using model M is:*

$$|C| = |M| + |D \mid M|,$$

where $|M|$ is the model size (the cost of transmitting the event space structure and count table) and $|D \mid M|$ is the data given the model (the residual after prediction).

In UM terms:

- $|M| = |I| \times |O| \times \lceil \log_2 N \rceil$ *bits (the count table).*
- $|D \mid M| = N \times bpc$ *bits (the arithmetic-coded residual).*

Corollary 17 (MDL interpretation)**.** *Minimizing the total compressed size $|C| = |M| + |D \mid M|$ is the Minimum Description Length (MDL) principle. The optimal event space balances model complexity against prediction accuracy.*

For small data (N small), simple models win ($|M|$ dominates). For large data (N large), complex models win ($|D \mid M|$ dominates). The crossover point determines the "right" event space granularity for a given data size.

Example 18 (Byte KN models at different scales)**.** *From scale_kn.c experiments:*

Data size N	*Best KN order*	*bpc*
10^3	*2*	*4.51*
10^4	*3*	*3.42*
10^5	*4*	*2.89*
10^6	*5*	*2.56*
10^7	*5*	*2.29*
10^8	*6*	*2.00*

The optimal order increases with data size, exactly as MDL predicts: more data supports more complex models (longer contexts).

9 Duality Theorem

Theorem 19 (Compression–prediction duality). *For the UM on event space $E = I \times O$:*

1. ***Prediction → compression:*** *Any predictor $\hat{P}$ defines a compressor with rate $L(\hat{P}, D)$ bpc via arithmetic coding.*
2. ***Compression → prediction:*** *Any compressor with rate r bpc defines a predictor with cross-entropy at most $r + o(1)$ via the decompressor's implicit model.*
3. ***Optimal equivalence:*** *The optimal predictor and the optimal compressor achieve the same rate: the entropy rate h.*

Proof. 1. Arithmetic coding achieves rate $L + O(1/N)$.

2. Any compressor that achieves rate r on all data from a source must implicitly model the source with cross-entropy at most r (by Shannon's source coding theorem).
3. The optimal predictor achieves $L = h$ (Theorem **??**). The optimal compressor achieves $r = h$ (source coding theorem). Both converge to the entropy rate.

□

Remark 20. *The duality theorem says: there is ONE mathematical object (the source's probability distribution) that underlies both compression and prediction. The UM's count table is a finite-sample estimate of this object. The forward pass extracts predictions; arithmetic coding extracts compressed bits. Both are views of the same information.*

10 Implications for the Hutter Prize

Proposition 21 (Path to winning). *To achieve k bpc on enwik8 ($N = 10^8$), the UM needs:*

1. *Event spaces capturing $8 - k$ bits of MI per position.*
2. *Sufficient data for the count table to converge.*
3. *An arithmetic coder wrapping the predictor.*

Current status (February 2026):

- *Byte KN-6 at 10^8: 2.00 bpc. Captures 6.00 bits/position.*
- *Sat-RNN at 10^3: 0.079 bpc. Captures 7.92 bits/position (but only on 1024 bytes of training data).*
- *Skip-8-gram at 10^3: 0.043 bpc. Captures 7.96 bits/position.*

The gap between 2.00 bpc (byte KN at scale) and 0.043 bpc (skip-8-gram at 10^3) is the scaling challenge: the skip-gram's MI is available in the data, but the count table grows combinatorially with context length.

Remark 22. *The compression–prediction duality frames the Hutter Prize purely as a prediction problem: achieve the lowest possible cross-entropy on enwik8. The compressor is mechanical (arithmetic coding). All the intelligence is in the predictor (the UM). And the predictor is entirely determined by the event space and the counting function. The Hutter Prize is won by the right event spaces.*

References

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

[2] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[3] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[4] Claude and MJC. *The Carrier Signal Problem*. Hutter archive, 12 Feb 2026.