Computational Information: A Generalization of Shannon Entropy

Asvin G.

January 12, 2026

Abstract

We present a framework for information that generalizes Shannon entropy along two orthogonal axes: (1) restricting the class of distributions available to model the truth, and (2) assigning computational costs to models. Classical entropy emerges as a special case, but with an interesting twist: the log-sum inequality, which classically appears in proving subadditivity, is factored into the optimization step that identifies entropy as a minimum. This clarifies the structure of classical information theory and provides a natural generalization to computationally bounded observers.

This is an alternate interpretation of [1], from a different vantage point. The motivations are the same however: to define entropy from an observer-dependent point of view of bounded computational power. Of course, this requires giving a definition of an observer in the first place, which we propose. The set up is as follows.

1 Subjective entropy

We consider the question of trying to approximate a noisy function. Let (X, μ_X) be set of possible pasts and Y the set of possible outcomes (both of which we assume to be discrete and finite for simplicity). Our target noisy function is $p : X \to \mathcal{P}(Y)$ where $p_x := p(x)$ is a probability distribution on Y.

Our observer $\mathcal{O}$ will consist of a space of "programs" $u : X \to \mathcal{P}(Y)$ along with an associated cost function $K(u) \in \mathbb{R}_{\geq 0}$. The observer attempts to find the best approximation to p by minimizing the "cost to describe p".

Our main motivation comes from machine learning. For instance, a LLM provides such an example. We take $Y = \Sigma$ to be a finite alphabet (the tokens), $X = \Sigma^*$ to be finite words on Σ and $p : \Sigma^* \to \mathcal{P}(\Sigma)$ to be the distribution on natural text. The transformer architecture with all parameters fixed then provides us with our observer $\mathcal{L}$ - a program $u \in \mathcal{L}$ corresponds to a specific choice of parameters (i.e., a trained network) while $K(u)$ can be taken to be the Kolmogorov complexity of u or even the cost to find u through gradient descent.

Definition 1 (Subjective entropy)**.** *The subjective entropy of p with respect to $\mathcal{O}$ is defined by*

$$\mathbb{M}_{\mathcal{O}}(p) := \min_u \{K(u) + H(p||u)$$

where $H(p||u) = \mathbb{E}_{x \sim \mu_X} H(p_x \,\|\, u_x) = \mathbb{E}_{x \sim \mu_X} \mathbb{E}_{y \sim p_x}[-\log u_x(y)]$ is the cross-entropy.

The cross-entropy has a coding interpretation: $H(p_x \,\|\, u_x)$ is the expected number of bits to encode a sample from p_x using a code optimized for u_x. Thus $\mathbb{M}_{\mathcal{O}}(p)$ is the minimum total description length: program cost plus expected encoding cost.

At the optimum u^*, we decompose:

$$\mathbf{S}_{\mathcal{O}}(p) = K(u^*) \qquad \text{(structural information—cost of the program)} \tag{1}$$
$$\mathbf{H}_{\mathcal{O}}(p) = H(p \,\|\, u^*) \qquad \text{(residual entropy—encoding cost given program)} \tag{2}$$

If our noisy function is noiseless, i.e., p_x is always a point mass, then the residual entropy is simply a measure of the accuracy of our approximation. If we take $\mathcal{O}$ to be the set of programs under some classic resource bound (for instance polynomial time programs), then we are in the land of classical complexity theory.

On the other hand, if we take $X = \{*\}$ to be a singleton, then our programs are simply distributions. If moreover, we assume that $K \equiv 0$, then we obtain a generalization of Shannon entropy.

Definition 2 (Shannon Observer). *The Shannon observer $\mathcal{S}$ is defined by $X = \{*\}, K \equiv 0$ and $\mathcal{S} = \mathcal{P}(Y)$ is the set of all distributions on Y. In this case,*

$$\mathbb{M}_{\mathcal{S}}(p) = \min_{u \in P(Y)} H(p||u) = \min_{u} H(p) + D_{KL}(p||u) = H(p)$$

since $D_{KL}(p||u) \geq 0$ with equality precisely when $u = p$.

By the same argument, we see that $H_{\mathcal{O}}(p) \geq H(p)$ for any observer $\mathcal{O}$.

This connection to Shannon entropy is the place where the convexity of the logarithm is crucial. As we will see, subjective entropy satisfies similar properties to Shannon entropy, but the proofs are essentially "free".

To illustrate the idea, let us look at the *uniform observer* $\mathcal{U} = \{$ uniform distribution on $Y\}$ with $K \equiv 0$ and $X = \{*\}$ again. For any distribution p, we have

$$\mathbb{M}_{\mathcal{U}}(p) = \mathbb{E}_p[-\log(1/|Y|)] = \log|Y|.$$

This observer cannot exploit any structure in p.

Autoregressive Models and the Shannon Limit

The most natural setting for our framework is autoregressive modeling, where the idealizations required to recover Shannon entropy become explicit.

Definition 3 (Autoregressive Model). *An* autoregressive model *over alphabet Σ is a program $u : X \to \mathcal{P}(Y)$ where $X = \Sigma^*$ (histories) and $Y = \Sigma$ (next token). Given history $x = (x_1, \ldots, x_{n-1})$, the program outputs a distribution $u_x \in \mathcal{P}(\Sigma)$ over the next symbol. This induces a joint distribution on sequences:*

$$p_u(x_1, \ldots, x_n) = \prod_{i=1}^{n} u_{x_{<i}}(x_i)$$

In our framework, an autoregressive model is precisely a program $u : X \to \mathcal{P}(Y)$. The observer $\mathcal{O}$ specifies which programs are available and their costs.

Proposition 1 (Subjective Entropy of Next-Token Prediction). *Let $p : X \to \mathcal{P}(Y)$ be the true next-token distribution given history (where $X = \Sigma^*$, $Y = \Sigma$). Then:*

$$\mathbb{M}_{\mathcal{O}}(p) = \min_{u \in \mathcal{O}} \{K(u) + \mathbb{E}_{x \sim \mu_X} \mathbb{E}_{y \sim p_x}[-\log u_x(y)]\}$$

The second term is exactly the per-token cross-entropy loss familiar from language modeling.

Classical Shannon entropy emerges under three idealizations:

1. **All conditionals available:** $\mathcal{O} = \{\text{all functions } X \to \mathcal{P}(Y)\}$
2. **Zero model cost:** $K \equiv 0$
3. **Ergodic/stationary limit:** The entropy rate exists

Theorem 1 (Shannon Entropy as a Limiting Case)**.** *Under idealizations (1) and (2), for each position n:*

$$\mathbb{M}_{\mathcal{S}}(p_n | x_{<n}) = H(Y_n | X_{<n})$$

where $\mathcal{S}$ denotes the Shannon observer. Under all three idealizations, for a stationary ergodic source:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{M}_{\mathcal{S}}(p_i | x_{<i}) = H(Y)$$

where $H(Y)$ is the Shannon entropy rate.

Proof. Under idealizations (1) and (2), the optimal program at position n is $u_x^* = p_x$, the true conditional distribution. The subjective entropy becomes:

$$\mathbb{M}_{\mathcal{S}}(p_n | x_{<n}) = 0 + H(p_n \,\|\, p_n) = H(Y_n | X_{<n})$$

The entropy rate formula follows from the standard ergodic theorem. □

Remark 1 (The Three Gaps from Shannon)**.** *Each idealization, when relaxed, creates a gap from classical entropy:*

- ***Restricted programs:*** *If $\mathcal{O} \subsetneq \{$all conditionals$\}$, we have $\mathbb{M}_{\mathcal{O}}(p) > H(p)$. For instance, if $\mathcal{O}$ contains only k-th order Markov models, the gap measures how much structure in p exceeds k-th order dependencies.*
- ***Positive program cost:*** *If $K(u) > 0$, there is a trade-off between model complexity and fit. This is the MDL (Minimum Description Length) perspective: the optimal u^* may* not *equal p, instead preferring a simpler program that accepts higher cross-entropy.*
- ***Finite sequences:*** *Before the asymptotic limit, there are edge effects. The first few tokens have less history, making prediction harder.*

Remark 2 (Connection to Language Modeling)**.** *Modern language models (transformers) provide a concrete instance with:*

- *$\mathcal{O} = \{$transformer architectures with $\leq N$ parameters$\}$*
- *$K(u) =$ training cost or parameter count (in various metrics)*

The gap $\mathbb{M}_{\mathcal{O}}(p_{\text{natural language}}) - H(p_{\text{natural language}})$ measures how far current architectures are from the "true" entropy of natural language—a quantity that Shannon entropy assumes is achievable for free.

2 Properties of subjective entropy

We now establish that $\mathbb{M}_{\mathcal{O}}$ satisfies the key properties of entropy (subadditivity, chain rule) in full generality. The proofs are remarkably simple—they just exhibit feasible solutions. The "hard work" (log inequalities) only appears when we specialize to the Shannon observer. In this section, we will want to vary the spaces X, Y, in particular to take products. For simplicity of notation, we consider the programs $u : X \to \mathcal{P}(Y)$ to be implicitly typed with source X and target Y, and allow for the observer $\mathcal{O}$ to have programs $u : X \to \mathcal{P}(Y)$ with X, Y varying.

Definition 4 (Product of Programs)**.** *For programs $u_i : X_i \to \mathcal{P}(Y_i)$, $i = 1, 2$, we define $u_1 \otimes u_2 : X_1 \times X_2 \to \mathcal{P}(Y_1 \times Y_2)$ by*

$$(u_1 \otimes u_2)_{(x_1,x_2)}(y_1, y_2) = u_{1,x_1}(y_1)u_{2,x_2}(y_2).$$

Definition 5 (Product-Closed Observer)**.** *An observer $\mathcal{O}$ is* product-closed with overhead c *if for all $u_i : X_i \to \mathcal{P}(Y_i)$:*

1. $u_1 \otimes u_2 \in \mathcal{O}$
2. $K(u_1 \otimes u_2) \leq K(u_1) + K(u_2) + c$

The Shannon observer is product-closed with $c = 0$.

Conversely, given $p : X_1 \times X_2 \to \mathcal{P}(Y_1 \times Y_2)$, we define the marginals by

Definition 6 (Marginals)**.** *We marginalize over the second variable by pushing forward p_{x_1,x_2} along $Y_1 \times Y_2 \to Y_2$ and integrating out the x_2. Explicitly,*

$$p_1 : X_1 \to \mathcal{P}(Y_1); p_{1,x_1}(y_1) = \mathbb{E}_{x_2 \in X_2} \sum_{y_2 \in Y_2} p_{x_1,x_2}(y_1, y_2)$$

and similarly for p_2.

With these two definitions, we recover the classical sub-additivity of Shannon entropy for subjective entropy.

Theorem 2 (Subadditivity)**.** *Let $\mathcal{O}$ be product-closed with overhead c. For any joint distribution $p : X_1 \times X_2 \to \mathcal{P}(Y_1 \times Y_2)$ with marginals p_1, p_2:*

$$\mathbb{M}_{\mathcal{O}}(p) \leq \mathbb{M}_{\mathcal{O}}(p_1) + \mathbb{M}_{\mathcal{O}}(p_2) + c \tag{3}$$

Proof. Let u_1^*, u_2^* be optimal programs for p_1, p_2. Consider the product program $u_1^* \otimes u_2^*$. The key calculation uses only marginalization (no log inequalities):

$$\begin{aligned} H(p \,\|\, u_1^* \otimes u_2^*) &= \mathbb{E}_{x_1,x_2 \in X_1 \times X_2} \mathbb{E}_{(y_1,y_2) \sim p}[-\log u_1^*(y_1) - \log u_2^*(y_2)] && (4) \\ &= \mathbb{E}_{x_1 \in X_1} \mathbb{E}_{y_1 \sim p_1}[-\log u_{1,x_1}^*(y_1)] + \mathbb{E}_{x_2 \in X_2} \mathbb{E}_{y_2 \sim p_2}[-\log u_{2,x_2}^*(y_2)] && (5) \\ &= H(p_1 \,\|\, u_1^*) + H(p_2 \,\|\, u_2^*) && (6) \end{aligned}$$

The equalities all follow simply from the definitions of marginalization and product measures without using any specific properties of the log function. As a consequence,

$$\begin{aligned} \mathbb{M}_{\mathcal{O}}(p) &\leq K(u_1^* \otimes u_2^*) + H(p \,\|\, u_1^* \otimes u_2^*) && (7) \\ &\leq [K(u_1^*) + K(u_2^*) + c] + [H(p_1 \,\|\, u_1^*) + H(p_2 \,\|\, u_2^*)] && (8) \\ &= \mathbb{M}_{\mathcal{O}}(p_1) + \mathbb{M}_{\mathcal{O}}(p_2) + c && \square \end{aligned}$$

When we specialize to the Shannon observer, we recover classical subadditivity $H(p) \leq H(p_1) + H(p_2)$.

When $\mathcal{O} = \mathcal{S}$ (so $K \equiv 0$, $c = 0$):

$$H(p) = \mathbb{M}_{\mathcal{S}}(p) \leq \mathbb{M}_{\mathcal{S}}(p_1) + \mathbb{M}_{\mathcal{S}}(p_2) = H(p_1) + H(p_2) \tag{9}$$

Thus, our framework *factors* the classical proof of subadditivity:

1. **Structural step** (Theorem 2): $\mathbb{M}(p) \leq \mathbb{M}(p_1) + \mathbb{M}(p_2)$—easy, no log inequalities

2. **Optimization step**: $\mathbb{M}_{\mathcal{S}}(p) = H(p)$—uses $D_{KL} \geq 0$

The classical proof combines these steps, obscuring the factorization.

Similarly, the chain rule holds in the general framework. The idea is that we can decompose the task of approximating p into two steps: first approximate a "summary statistic" of p, then approximate p given that statistic.

Setup for the Chain Rule

Let $p : X \to \mathcal{P}(Y)$ be our target and $\pi : Y \to Z$ a projection (surjective map) extracting some information from Y. We think of Z as a coarsening or summary of Y.

Definition 7 (Pushforward). *The pushforward $\pi_* p : X \to \mathcal{P}(Z)$ is defined by*

$$(\pi_* p)_x(z) = \sum_{y:\pi(y)=z} p_x(y) = p_x(\pi^{-1}(z)).$$

This is the distribution on Z induced by first sampling $y \sim p_x$, then computing $z = \pi(y)$.

Definition 8 (Conditional Distribution). *The conditional $p|_\pi : X \times Z \to \mathcal{P}(Y)$ is defined by*

$$(p|_\pi)_{x,z}(y) = \begin{cases} \frac{p_x(y)}{(\pi_* p)_x(z)} & \textit{if } \pi(y) = z \\ 0 & \textit{otherwise} \end{cases}$$

This is the distribution on Y given that we know x and have observed $z = \pi(y)$. Note that $(p|_\pi)_{x,z}$ is supported on the fiber $\pi^{-1}(z)$.

Definition 9 (Conditional Subjective Entropy). *The conditional subjective entropy of p given π is:*

$$\mathbb{M}_{\mathcal{O}}(p|\pi) := \min_{u:X\times Z\to\mathcal{P}(Y)} \left\{ K(u) + \mathbb{E}_{x\sim\mu_X} \mathbb{E}_{z\sim(\pi_* p)_x} \left[H((p|_\pi)_{x,z} \parallel u_{x,z}) \right] \right\}$$

where the inner expectation is over z distributed according to the pushforward $(\pi_* p)_x$.

The key insight is that approximating p can be decomposed into: (1) approximating $\pi_* p$ (the summary), and (2) approximating $p|_\pi$ (the detail given the summary).

Definition 10 (Composition of Programs). *Given $u : X \to \mathcal{P}(Z)$ (a program for the summary) and $v : X \times Z \to \mathcal{P}(Y)$ (a conditional program for Y given Z), we define the composed program $u \circ_\pi v : X \to \mathcal{P}(Y)$ by*

$$(u \circ_\pi v)_x(y) = u_x(\pi(y)) \cdot v_{x,\pi(y)}(y).$$

Definition 11 (Composition-Closed Observer). *An observer $\mathcal{O}$ is* composition-closed with overhead c *(with respect to $\pi : Y \to Z$) if for all programs $u : X \to \mathcal{P}(Z)$ and $v : X \times Z \to \mathcal{P}(Y)$ in $\mathcal{O}$:*

1. $u \circ_\pi v \in \mathcal{O}$

2. $K(u \circ_\pi v) \leq K(u) + K(v) + c$

Theorem 3 (Chain Rule)**.** *Let $\mathcal{O}$ be composition-closed with overhead c with respect to $\pi : Y \to Z$. Then:*

$$\mathbb{M}_{\mathcal{O}}(p) \leq \mathbb{M}_{\mathcal{O}}(\pi_* p) + \mathbb{M}_{\mathcal{O}}(p|\pi) + c$$

Proof. Let u^* be optimal for $\pi_* p$ and v^* be optimal for $p|_\pi$. Consider the composed program $u^* \circ_\pi v^*$.

Key calculation: The cross-entropy factors:

$$\begin{aligned} H(p \,\|\, u^* \circ_\pi v^*) &= \mathbb{E}_x \mathbb{E}_{y \sim p_x} \left[-\log u^*_x(\pi(y)) - \log v^*_{x,\pi(y)}(y) \right] & (10) \\ &= \mathbb{E}_x \mathbb{E}_{y \sim p_x} \left[-\log u^*_x(\pi(y)) \right] + \mathbb{E}_x \mathbb{E}_{y \sim p_x} \left[-\log v^*_{x,\pi(y)}(y) \right] & (11) \end{aligned}$$

For the first term, since $z = \pi(y)$ and $y \sim p_x$ implies $z \sim (\pi_* p)_x$:

$$\mathbb{E}_x \mathbb{E}_{y \sim p_x} \left[-\log u^*_x(\pi(y)) \right] = \mathbb{E}_x \mathbb{E}_{z \sim (\pi_* p)_x} \left[-\log u^*_x(z) \right] = H(\pi_* p \,\|\, u^*)$$

For the second term, we condition on $z = \pi(y)$:

$$\mathbb{E}_x \mathbb{E}_{y \sim p_x} \left[-\log v^*_{x,\pi(y)}(y) \right] = \mathbb{E}_x \mathbb{E}_{z \sim (\pi_* p)_x} \mathbb{E}_{y \sim (p|_\pi)_{x,z}} \left[-\log v^*_{x,z}(y) \right] = H(p|_\pi \,\|\, v^*)$$

Therefore:

$$\begin{aligned} \mathbb{M}_{\mathcal{O}}(p) &\leq K(u^* \circ_\pi v^*) + H(p \,\|\, u^* \circ_\pi v^*) & (12) \\ &\leq [K(u^*) + K(v^*) + c] + [H(\pi_* p \,\|\, u^*) + H(p|_\pi \,\|\, v^*)] & (13) \\ &= \mathbb{M}_{\mathcal{O}}(\pi_* p) + \mathbb{M}_{\mathcal{O}}(p|\pi) + c & \square \end{aligned}$$

When is the Chain Rule an Equality?

The chain rule gives an upper bound. For a matching lower bound, we need a converse condition.

Definition 12 (Decomposition of Programs)**.** *Every program $u : X \to \mathcal{P}(Y)$ decomposes through π as $u = \bar{u} \circ_\pi u|_\pi$ where:*

- *$\bar{u} : X \to \mathcal{P}(Z)$ is the marginal: $\bar{u}_x = \pi_*(u_x)$*
- *$u|_\pi : X \times Z \to \mathcal{P}(Y)$ is the conditional: $(u|_\pi)_{x,z}(y) = u_x(y)/\bar{u}_x(z)$ for $\pi(y) = z$*

This is simply the factorization $u_x(y) = \bar{u}_x(\pi(y)) \cdot (u|_\pi)_{x,\pi(y)}(y)$.

Definition 13 (Decomposition-Closed Observer)**.** *An observer $\mathcal{O}$ is* decomposition-closed with overhead c' *(with respect to $\pi : Y \to Z$) if for every program $u : X \to \mathcal{P}(Y)$ in $\mathcal{O}$:*

1. $\bar{u} \in \mathcal{O}$ *and* $u|_\pi \in \mathcal{O}$

2. $K(u) \geq K(\bar{u}) + K(u|_\pi) - c'$

The second condition says that decomposing a program does not make it cheaper (up to overhead c'). This is the reverse of composition-closed.

Theorem 4 (Chain Rule Equality). *Let $\mathcal{O}$ be both composition-closed with overhead c and decomposition-closed with overhead c' (with respect to $\pi : Y \to Z$). Then:*

$$\mathbb{M}_{\mathcal{O}}(\pi_* p) + \mathbb{M}_{\mathcal{O}}(p|\pi) - c' \leq \mathbb{M}_{\mathcal{O}}(p) \leq \mathbb{M}_{\mathcal{O}}(\pi_* p) + \mathbb{M}_{\mathcal{O}}(p|\pi) + c$$

Proof. The upper bound is Theorem 3. For the lower bound, let u^* be optimal for p. By decomposition-closure, $\bar{u}^* \in \mathcal{O}$ and $u^*|_\pi \in \mathcal{O}$.

The cross-entropy factors (same calculation as before):

$$H(p \,\|\, u^*) = H(\pi_* p \,\|\, \bar{u}^*) + H(p|_\pi \,\|\, u^*|_\pi)$$

Therefore:

$$\begin{aligned}
\mathbb{M}_{\mathcal{O}}(p) &= K(u^*) + H(p \,\|\, u^*) && (14)\\
&\geq [K(\bar{u}^*) + K(u^*|_\pi) - c'] + [H(\pi_* p \,\|\, \bar{u}^*) + H(p|_\pi \,\|\, u^*|_\pi)] && (15)\\
&= [K(\bar{u}^*) + H(\pi_* p \,\|\, \bar{u}^*)] + [K(u^*|_\pi) + H(p|_\pi \,\|\, u^*|_\pi)] - c' && (16)\\
&\geq \mathbb{M}_{\mathcal{O}}(\pi_* p) + \mathbb{M}_{\mathcal{O}}(p|\pi) - c' && (17)
\end{aligned}$$

where the last inequality uses that $\bar{u}^*$ is feasible (not necessarily optimal) for $\pi_* p$, and similarly for $u^*|_\pi$. □

Remark 3 (When $c = c' = 0$: Exact Equality). *If $\mathcal{O}$ is both composition-closed and decomposition-closed with $c = c' = 0$, then:*

$$\mathbb{M}_{\mathcal{O}}(p) = \mathbb{M}_{\mathcal{O}}(\pi_* p) + \mathbb{M}_{\mathcal{O}}(p|\pi)$$

This is the case for the Shannon observer, where $K \equiv 0$ trivially satisfies both conditions.

Remark 4 (Classical Chain Rule). *In the Shannon case ($\mathcal{O} = \mathcal{S}$, $K \equiv 0$), this becomes:*

$$H(Y) = H(\pi(Y)) + H(Y|\pi(Y))$$

which is the classical chain rule as an equality.

Remark 5 (Interpretation). *The chain rule says: to describe Y, we can first describe the summary $\pi(Y)$, then describe Y given the summary.*

Equality holds when there is no "compression advantage" to representing p jointly versus as (summary, conditional). If knowing the summary helps compress the conditional description (shared structure), you get strict inequality $\mathbb{M}(p) < \mathbb{M}(\pi_* p) + \mathbb{M}(p|\pi)$.

3 Connection to Epiplexity

Our framework connects to the recent notion of *epiplexity* [1].

Definition 14 (Epiplexity, after Finzi et al.). *Fix a time bound T. Let $\mathcal{O}_T$ be time-bounded probabilistic programs. The **epiplexity** and **time-bounded entropy** of distribution p are:*

$$\begin{aligned}
\mathrm{S}_T(p) &= |u^*| && \textit{(program length of optimal program)} && (18)\\
\mathrm{H}_T(p) &= \mathbb{E}_{y\sim p}[-\log u^*(y)] && \textit{(cross-entropy with optimal program)} && (19)
\end{aligned}$$

where $u^* = \arg\min_{u \in \mathcal{O}_T}\{|u| + \mathbb{E}_{y\sim p}[-\log u(y)]\}$.

This is precisely our framework with:

- $\mathcal{O} = \mathcal{O}_T$ (time-bounded programs)
- $K(u) = |u|$ (program length)

Our contribution is to clarify the structure: epiplexity combines *two* generalizations (restricted programs and positive costs) that can be studied independently.

4 Shearer's Inequality for Subjective Entropy

We now generalize Shearer's inequality, a powerful refinement of subadditivity, to subjective entropy.

Classical Shearer's Inequality

Let $Y = Y_1 \times \cdots \times Y_n$ and let $S_1, \ldots, S_m \subseteq [n]$ be subsets such that each index $i \in [n]$ appears in at least k of the subsets. For $S \subseteq [n]$, write $Y_S = \prod_{i \in S} Y_i$ and let $\pi_S : Y \to Y_S$ be the projection.

Theorem 5 (Classical Shearer)**.** *For any distribution p on Y:*

$$H(Y_1, \ldots, Y_n) \leq \frac{1}{k} \sum_{j=1}^{m} H(Y_{S_j})$$

Special cases include:

- **Subadditivity:** $S_i = \{i\}$, $k = 1$ gives $H(Y) \leq \sum_i H(Y_i)$
- **Han's inequality:** $S_i = [n] \setminus \{i\}$, $k = n - 1$ gives $H(Y) \leq \frac{1}{n-1} \sum_i H(Y_{[n]\setminus\{i\}})$

The Fractional Product Construction

To generalize Shearer, we need to construct a program for p on Y from programs u_j on the marginals Y_{S_j}.

Definition 15 (Fractional Product)**.** *Given programs $u_j : X \to \mathcal{P}(Y_{S_j})$ for $j = 1, \ldots, m$, the* fractional product with exponent $1/k$ *is:*

$$u_x^{\otimes}(y) = \frac{1}{Z_x} \prod_{j=1}^{m} u_{j,x}(y_{S_j})^{1/k}$$

where $Z_x = \sum_{y \in Y} \prod_{j=1}^{m} u_{j,x}(y_{S_j})^{1/k}$ is the normalization constant.

Proposition 2 (Cross-Entropy of Fractional Product)**.**

$$H(p \,\|\, u^{\otimes}) = \mathbb{E}_x[\log Z_x] + \frac{1}{k} \sum_{j=1}^{m} H(\pi_{S_j,*}p \,\|\, u_j)$$

Proof.

$$\begin{aligned}
H(p \,\|\, u^{\otimes}) &= \mathbb{E}_x \mathbb{E}_{y \sim p_x} \left[-\log u_x^{\otimes}(y) \right] && (20)\\
&= \mathbb{E}_x \mathbb{E}_{y \sim p_x} \left[\log Z_x - \frac{1}{k} \sum_j \log u_{j,x}(y_{S_j}) \right] && (21)\\
&= \mathbb{E}_x[\log Z_x] + \frac{1}{k} \sum_j \mathbb{E}_x \mathbb{E}_{y \sim p_x} \left[-\log u_{j,x}(y_{S_j}) \right] && (22)\\
&= \mathbb{E}_x[\log Z_x] + \frac{1}{k} \sum_j H(\pi_{S_j,*}p \,\|\, u_j) && \square
\end{aligned}$$

The key observation is that Shearer's inequality holds if $\mathbb{E}_x[\log Z_x]$ is bounded.

Shearer-Closed Observers

Definition 16 (Shearer-Closed Observer)**.** *An observer $\mathcal{O}$ is* Shearer-closed *with respect to cover $(S_1, \ldots, S_m)$ with coverage k, overhead c, and normalization bound c' if for any programs $u_j : X \to \mathcal{P}(Y_{S_j})$ in $\mathcal{O}$:*

1. *The fractional product $u^{\otimes} \in \mathcal{O}$*
2. $K(u^{\otimes}) \leq \frac{1}{k} \sum_{j=1}^{m} K(u_j) + c$
3. $\mathbb{E}_x[\log Z_x] \leq c'$ *(normalization bound)*

Condition (3) is where the "log inequality" hides. For the Shannon observer with optimal choices $u_j = \pi_{S_j,*}p$, the classical Shearer inequality implies $\log Z \leq 0$.

Theorem 6 (Shearer's Inequality for Subjective Entropy)**.** *Let $\mathcal{O}$ be Shearer-closed with respect to $(S_1, \ldots, S_m)$ with coverage k, overhead c, and normalization bound c'. Then for any $p : X \to \mathcal{P}(Y)$:*

$$\mathbb{M}_{\mathcal{O}}(p) \leq \frac{1}{k} \sum_{j=1}^{m} \mathbb{M}_{\mathcal{O}}(\pi_{S_j,*}p) + c + c'$$

Proof. Let u_j^* be optimal for $\pi_{S_j,*}p$. Consider the fractional product $u^{\otimes}$ built from these.

By the cross-entropy formula and the Shearer-closed conditions:

$$\begin{aligned}
\mathbb{M}_{\mathcal{O}}(p) &\leq K(u^{\otimes}) + H(p \,\|\, u^{\otimes}) && (23)\\
&\leq \left[\frac{1}{k} \sum_j K(u_j^*) + c\right] + \left[\mathbb{E}_x[\log Z_x] + \frac{1}{k} \sum_j H(\pi_{S_j,*}p \,\|\, u_j^*)\right] && (24)\\
&\leq \frac{1}{k} \sum_j \left[K(u_j^*) + H(\pi_{S_j,*}p \,\|\, u_j^*)\right] + c + c' && (25)\\
&= \frac{1}{k} \sum_j \mathbb{M}_{\mathcal{O}}(\pi_{S_j,*}p) + c + c'
\end{aligned}$$

□

The Shannon Observer Satisfies Shearer

Proposition 3. *The Shannon observer $\mathcal{S}$ is Shearer-closed with $c = c' = 0$ for any cover with coverage k.*

Proof. Conditions (1) and (2) are trivial since $K \equiv 0$ and all distributions are in $\mathcal{S}$.

For condition (3), we need $\log Z \leq 0$ when $u_j = \pi_{S_j,*}p$ (the optimal choices). This is equivalent to showing:

$$\sum_y \prod_j p(y_{S_j})^{1/k} \leq 1$$

This is precisely the content of the classical Shearer inequality! Indeed, classical Shearer can be written as:

$$\sum_y p(y) \log p(y) \geq \frac{1}{k} \sum_j \sum_y p(y) \log p(y_{S_j})$$

which, by convexity arguments, implies the normalization bound.

Alternatively: by Hölder's inequality with exponents summing to k (each coordinate appears k times), we have $Z \leq 1$. □

Remark 6 (Where the Log Inequality Hides)**.** *Just as with subadditivity and the chain rule, the log inequality in Shearer appears in the* optimization/normalization step *(showing* $\log Z \leq 0$*) rather than in the* structural step *(the algebraic manipulation of cross-entropies).*

For general observers, we simply assume *the normalization bound as part of the Shearer-closed condition. This isolates exactly what property is needed beyond the algebraic structure.*

Remark 7 (Special Cases)**.**

- *For* $S_i = \{i\}$ *and* $k = 1$*: Shearer reduces to subadditivity, and the fractional product is just the ordinary product. The normalization bound* $Z \leq 1$ *is trivial (it's an equality).*
- *For* $m = 2$*,* $S_1 \cup S_2 = [n]$*,* $S_1 \cap S_2 = \emptyset$*,* $k = 1$*: This is exactly subadditivity, recovered as a special case.*

References

[1] M. Finzi, S. Qiu, Y. Jiang, P. Izmailov, J. Z. Kolter, and A. G. Wilson. From entropy to epiplexity: Rethinking information for computationally bounded intelligence. *arXiv:2601.03220*, 2025.