ω_∞: The Sky-Hook Problem

Hutter RNN Project

2026-01-31

1 The Universal Model

Recall the CMP model:

$$u = (E, T, P, f, \omega)$$

- E = event space (e.g., 256 bytes)
- T = thought vectors in $\mathbb{N}^{|E|}$
- P = patterns (transition weights)
- f = integration function
- ω = **prior over patterns**

The prior ω is where the magic happens. After seeing data D:

$$\omega_D = \text{posterior} = \frac{P(D|\omega_0) \cdot \omega_0}{P(D)}$$

But what is ω_0? The **prior before any data**.

2 The Sky-Hook

ω_∞ is the "sky-hook": the source of predictive power that doesn't come from data.

2.1 The Regress

1. Model learns patterns from data
2. But learning requires a prior ω_0
3. Where does ω_0 come from?
4. Either:
   - From more data (infinite regress)
   - From **something outside the system** (sky-hook)

2.2 Sources of ω_∞

In practice, the sky-hook is:

Source	Example
Evolution	Brain architecture, attention, memory
Human design	ES choice, RNN architecture, hyperparameters
Math	Minimum description length, Occam's razor
Physics	Locality, causality, symmetry

3 ω_∞ in Our System

3.1 What We Choose (Human Ingenuity)

1. **Event space factorization:** ES = {Digit, Punct, Vowel, Whitespace, Other}

   This is a *human choice*. We could have chosen differently.

2. **Architecture:** RNN with 512 hidden units, tanh activation

   Encodes locality bias, bounded memory, smooth functions.

3. **Training:** SGD with momentum, learning rate schedule

   Implicit regularization toward "simple" solutions.

3.2 Information Flow

$$\omega_\infty \xrightarrow{\text{human design}} \omega_0 \xrightarrow{\text{training}} \omega_D \xrightarrow{\text{compression}} \text{bits/char}$$

The question: *how much of final performance comes from ω_∞ vs. D?*

4 Measuring the Sky-Hook

4.1 Definition

The **sky-hook contribution** is:

$$\Delta_\omega = H(\text{data}|\text{uniform prior}) - H(\text{data}|\omega_\infty)$$

This is how much our prior assumptions help, independent of learning.

4.2 Empirical Approach

Compare:

1. Random weights (no ω_∞): ~ 8 bits/char
2. Untrained ES-augmented (ES structure only): $\sim$? bits/char

3. Trained on different data (transfer): $\sim$? bits/char
4. Trained on enwik9 (full): ~ 1.5 bits/char

The gap between 1 and 4 is *total* improvement.
The gap between 1 and 2 is *pure architecture* (ω_∞).

5 The ES Contribution to ω_∞

Our ES features encode:

$$\omega_\infty(\text{“vowels predict differently than consonants”}) > 0 \tag{1}$$
$$\omega_\infty(\text{“whitespace follows punctuation”}) > 0 \tag{2}$$
$$\omega_\infty(\text{“digits cluster”}) > 0 \tag{3}$$

These are **human-injected priors**. They came from:

- Linguistic knowledge (vowels vs consonants)
- Typography conventions (spacing)
- Domain knowledge (numbers group)

6 The Deep Question

Can we discover good ω_∞ automatically?

- Neural architecture search: Learn the architecture
- Meta-learning: Learn to learn (but from what meta-prior?)
- Compression: The “simplest” model that fits

But this just pushes the sky-hook up one level.

Solomonoff induction says: use the universal prior (Kolmogorov complexity). But this is uncomputable.

In practice: Human ingenuity is the sky-hook. We inject ω_∞ through our choices.

7 Implications

1. **Interpretability is about ω_∞:** We understand what we designed.
2. **The Hutter Prize rewards ω_∞:** Better priors = better compression.
3. **ES features are explicit ω_∞:** We can audit them.

The goal: Make ω_∞ visible, measurable, and improvable.