Discrete Codes from Continuous Substrates: A Minimal Channel-Capacity Simulation

Ian Todd
University of Sydney
itod2305@uni.sydney.edu.au

November 27, 2025

Abstract

We present a minimal simulation demonstrating that discrete symbolic codes emerge spontaneously when continuous data passes through a noisy bandwidth-limited channel. Using a simple autoencoder trained on a 1-D ring manifold embedded in 512-D, we show that neither the dimensional bottleneck nor channel noise alone produces discretization—both ingredients are required for the qualitative transition from continuous to discrete representation. We quantify this transition using the Adjusted Rand Index (semantic preservation) and participation ratio (effective dimensionality). The simulation provides a concrete, reproducible demonstration of how discrete codes can emerge from continuous substrates through information-theoretic constraints, with potential relevance to categorical perception, neural coding, and the emergence of digital information in biological systems.

1 Introduction

How do discrete symbols emerge from continuous physical substrates? This question spans cognitive science (how do brains form discrete concepts?), linguistics (how did discrete language emerge?), and origin-of-life research (how did digital genetic information arise from analog chemistry?).

We approach this question computationally, constructing a minimal simulation that exhibits spontaneous discretization. The goal is not to model any specific biological system, but to identify the *sufficient conditions* for code formation—the minimal ingredients that, when combined, produce discrete representations.

Our central finding is that two ingredients are required:

1. A **dimensional bottleneck**: Information must pass through a channel with fewer degrees of freedom than the input

2. **Channel noise**: The transmission is corrupted by stochastic perturbations

Neither alone produces discretization. The bottleneck without noise preserves continuous structure. Noise without a bottleneck leaves room for continuous codes. But together, they force the system into a regime where only discrete codes can reliably transmit information.

This paper describes what we simulated, how we simulated it, and what it might mean.

Table 1: Mapping between simulation components and physics concepts.

Simulation	**Physics Analogue**	**Mathematical Proxy**
Input data	Continuous manifold	S^1 with $O(2)$ symmetry
Bottleneck (k)	Degrees of freedom	Embedding dimension
Noise (σ)	Thermal fluctuations	Temperature-like parameter
Loss function	Energy-like objective	Effective potential to minimize
Code clustering	Symmetry reduction	$O(2) \to C_n$

2 What We Simulated

2.1 The Input: A Continuous Manifold

We need a continuous input space with clear topology. We chose a ring (S^1)—a 1-dimensional closed manifold—embedded in high-dimensional space:

```
# Generate 2000 points uniformly on a circle
theta = random(0, 2*pi, n_samples=2000)
ring_2d = [cos(theta), sin(theta)]

# Embed in 512 dimensions via random projection
projection = random_orthonormal(2, 512)
data_512d = ring_2d @ projection + noise(0.05)
```

The embedding noise (0.05) is small relative to the ring radius, so the intrinsic S^1 topology is preserved. The input is *continuous*—points can take any value along the ring. There are no discrete categories in the data itself.

For *evaluation only*, we partition the ring into 6 equal sectors and ask: do the learned codes preserve this structure? We treat these sectors as "semantic classes" and measure the Adjusted Rand Index (ARI) between k-means clusters of learned codes and ground-truth sectors.

2.2 The Channel: Encoder-Noise-Decoder

The core architecture is a sender-channel-receiver system:

```
# Encoder: compress 512D to k dimensions
code = encoder(input)  # code is k-dimensional

# Channel: add Gaussian noise
noisy_code = code + noise(sigma)

# Decoder: reconstruct 512D from noisy code
output = decoder(noisy_code)

# Training objective: minimize reconstruction error
loss = mean_squared_error(input, output)
```

The encoder and decoder are neural networks (3-layer MLPs with ReLU activations). The channel dimension k and noise level σ are the experimental variables.

The key insight is that the encoder must learn codes that are *robust to noise*. If two nearby codes get confused after noise corruption, the decoder cannot distinguish them, and reconstruction suffers. The only solution is to space codes apart—creating discrete clusters with "safety margins."

2.3 The Metrics

We measure two quantities:

Semantic Preservation (ARI): We cluster learned codes via k-means ($n = 6$) and compute the Adjusted Rand Index against the 6-sector ground truth. ARI = 1 means perfect cluster-label agreement; ARI = 0 means agreement no better than chance.

Effective Dimensionality (Participation Ratio):

$$\text{PR} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \tag{1}$$

where λ_i are eigenvalues of the code covariance matrix. PR = 1 for data collapsed to a line; PR $\approx d$ for variance spread evenly across d orthogonal directions. This serves as a coarse measure of effective embedding dimension.

3 The Control Experiment

Our main result comes from a simple control experiment. We ask: what *causes* the discretization?

Table 2: Control experiment: both bottleneck AND noise required for clustering.

Condition	Bottleneck (k)	Noise (σ)	Result
Bottleneck only	2	0	Continuous ring
Noise only	32	0.5	Continuous ring
Both	**2**	**0.5**	**Discrete clusters**

Figure 1 shows this visually. The top row tells a linear story:

1. **The Input**: A perfect continuous ring (color = angular position θ)
2. **Bottleneck Only**: Compressed to $k = 2$ with no noise—ring topology preserved
3. **Bottleneck + Noise**: Compressed to $k = 2$ with $\sigma = 0.5$—discrete clusters emerge

The bottom row shows that increasing k (with noise held constant) restores continuous representation. At $k = 32$, there is enough "room" in the channel for the ring to survive despite noise.

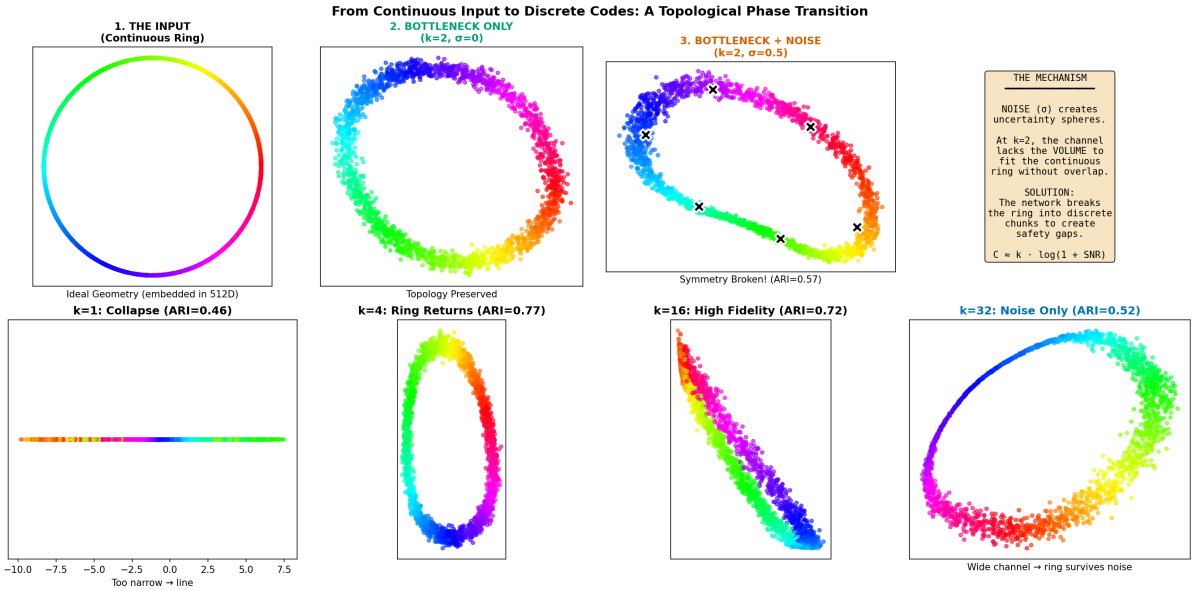

Figure 1: **Visual proof of the two-ingredient hypothesis.** Top row: (1) Continuous input ring, (2) bottleneck alone preserves topology, (3) bottleneck + noise → discrete clusters. Black X's mark learned centroids (k-means). Bottom row: As channel dimension k increases, ring structure returns despite noise ($\sigma = 0.5$ throughout).

4 The Transition

Figure 2 shows how semantic preservation varies with channel dimension k. There are three regimes:

1. $k = 1$**: Topological collapse.** The ring is forced onto a line. Adjacent points on the ring may map to distant points on the line. The S^1 topology is destroyed.

2. $k \approx k_c$**: Critical capacity.** Discrete codes emerge that optimally preserve semantic structure. The system "discovers" that a small number of well-separated clusters is the best way to use the limited channel.

3. $k \gg k_c$**: Continuous preservation.** With sufficient bandwidth, the ring can be transmitted continuously. Noise is present but doesn't force discretization.

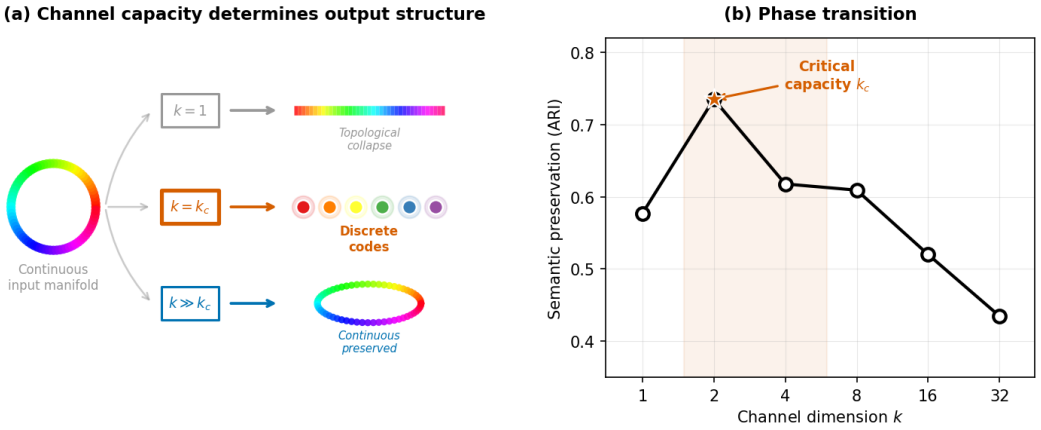

Figure 2: **Qualitative transition in code structure.** (a) Schematic showing three regimes. (b) Semantic preservation (ARI vs. 6-sector ground truth) as function of channel dimension k. Peak near $k_c \approx 2$. Noise $\sigma = 0.5$ throughout.

In the learned representation, there is a clear topology-changing transition: for $k = 1$ the ring maps to a line, at intermediate k to well-separated clusters, and for large k back to a continuous ring-like manifold. In the language of symmetry, this corresponds to a reduction from the continuous $O(2)$ rotational symmetry of the input ring to a discrete C_n point-group symmetry in code space.

We can define an order parameter for this transition as the silhouette score of the code distribution, or equivalently the ARI with respect to equal-sector partitions. This order parameter is near zero for $k = 1$ (collapsed) and $k \gg k_c$ (continuous), and peaks at intermediate k where discrete clustering is strongest.

Note: We use "transition" rather than "phase transition" advisedly. In a strict thermodynamic sense, phase transitions occur in the infinite-size limit where singularities appear in the free energy. Our finite neural network exhibits a sharp crossover or bifurcation, not a true thermodynamic phase transition. The symmetry-breaking language describes the *representation-level* structure, not a thermodynamic equilibrium.

5 The Mechanism: Capacity Constraints

Why does discretization occur? The intuition comes from **sphere packing** and **channel capacity**.

5.1 Geometric Picture

Channel noise smears each code point into an "uncertainty region" of characteristic scale $\sim \sigma$. If two codes are closer than this scale, their uncertainty regions overlap, and the decoder cannot reliably distinguish them.

The question becomes: how many well-separated codes can coexist in a k-dimensional channel? This is related to the classical sphere-packing problem [Conway and Sloane, 1999]. For fixed noise scale and code amplitude, only a finite number of non-overlapping regions fit in low-dimensional space.

5.2 Information-Theoretic Picture

The Shannon–Hartley theorem [Shannon, 1948] gives the capacity of an additive white Gaussian noise (AWGN) channel:

$$C \propto k \cdot \log_2(1 + \text{SNR}) \tag{2}$$

where k is the number of channel uses (dimensions) and SNR is the signal-to-noise ratio. The *logarithm* of the number of distinguishable codewords scales with this capacity.

We do not attempt a precise calibration here. The point is qualitative: for fixed k and σ, capacity is finite, so only a finite number of well-separated codes can coexist. For our parameters ($k = 2$, $\sigma = 0.5$), we empirically observe $\sim$6 clusters. We do not claim the theory uniquely predicts "6"—only that a small integer number of clusters is natural in this regime.

The network doesn't "know" about sphere packing or Shannon capacity; it discovers these constraints through gradient descent on reconstruction error.

6 What Gets Preserved

When compressing from 512 dimensions to k dimensions, what does the encoder choose to preserve?

Figure 3 shows that the encoder learns to preserve **high-variance directions**—essentially the top principal components of the input. The ring lives primarily in the first 2 PCs (we verified that PC1 $\approx \cos\theta$ and PC2 $\approx \sin\theta$ by plotting against θ), and these are what the encoder captures.

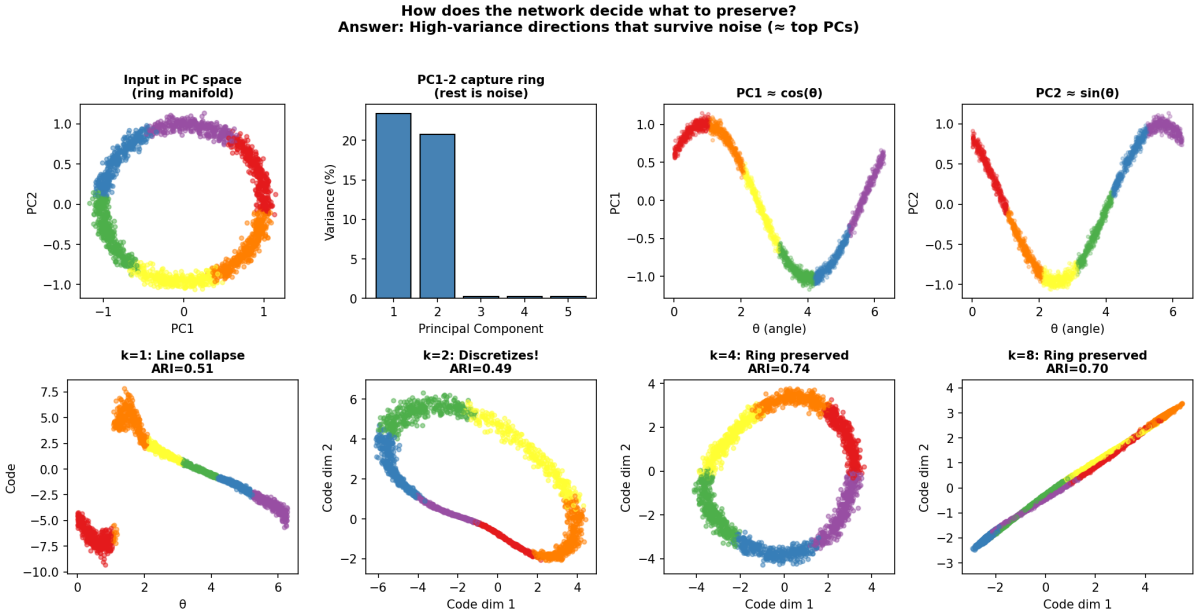

Figure 3: **What gets preserved.** Top: Input lives in first 2 PCs (intrinsic dimensionality ≈ 2). PC1-2 capture the ring; higher PCs are embedding noise. Bottom: Encoder learns to preserve these high-variance directions. At $k = 2$ with noise, discretization occurs.

This makes intuitive sense: high-variance directions carry the most information about input structure. Low-variance directions (the remaining 510 dimensions of embedding noise) are discarded as uninformative.

In the limit of a linear encoder/decoder and small noise, the optimum is PCA. Here the nonlinearity (ReLU) allows genuine topology changes—the "tearing" of the continuous ring into discrete clusters when capacity is insufficient.

7 Extension: Stochastic Resonance

Our framework provides a geometric interpretation of the **stochastic resonance** (SR) phenomenon—where adding noise can *improve* signal detection [Gammaitoni et al., 1998].

In the classical SR picture, a subthreshold periodic signal in a bistable potential is optimally detected at intermediate noise, which enables barrier crossing. Here, the "potential wells" are code basins learned by the autoencoder, and ambiguity arises from signals lying near decision boundaries.

Consider an ambiguous signal sitting exactly on the boundary between two code clusters. In low-noise conditions, the dynamics are confined to the low-dimensional manifold of learned codes—the signal is "trapped" on the decision boundary.

Adding noise allows the system to explore the ambient high-dimensional embedding space. The signal can fluctuate perpendicular to the boundary, enabling escape into the correct attractor basin by moving *around* the confusion region in high-D space rather than being stuck on a low-D saddle.

Figure 4 demonstrates this SR-like behavior:

- At zero noise, ambiguous signals are trapped on boundaries → poor decoding
- At intermediate noise, the system can escape to correct basins → peak accuracy
- At high noise, random walk dominates → decoding degrades

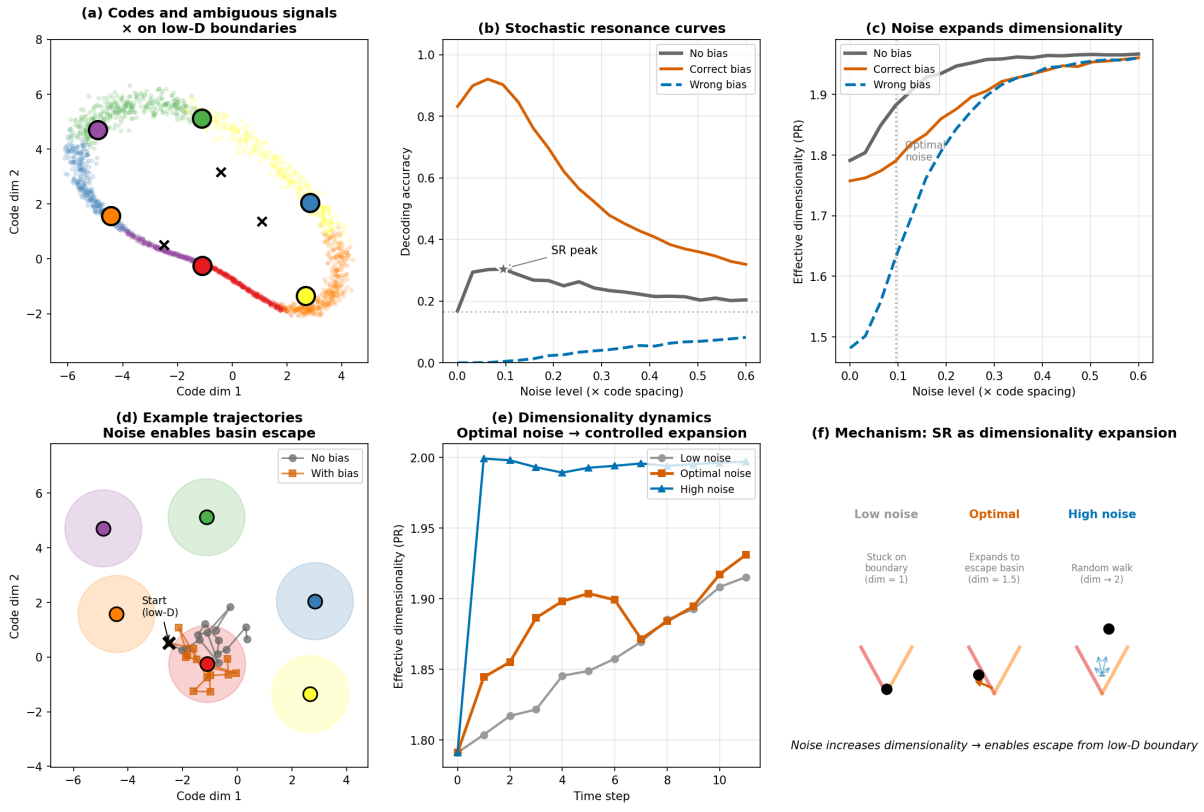

Figure 4: **SR-like behavior in code decoding.** (a) Learned codes and ambiguous test signals on basin boundaries. (b) Accuracy peaks at intermediate noise—the hallmark SR curve. (c) Effective dimensionality (PR) increases monotonically with noise. (d–f) Geometric interpretation: noise enables exploration of ambient space to escape low-D traps.

This reframes SR in geometric terms: noise provides access to extra degrees of freedom in the ambient embedding space, enabling escape from low-dimensional traps. Optimal performance occurs when noise is sufficient to escape boundaries but not so large as to cause random wandering.

8 Speculative Implications

The simulation is deliberately minimal and agnostic about biophysical details. The points below are *speculative connections*, not claims that this specific architecture underlies any particular system.

8.1 Categorical Perception

Humans perceive continuous stimuli as discrete categories—colors, phonemes, facial expressions [Harnad, 1987]. If sensory information must pass through bandwidth-limited neural pathways (due to metabolic constraints, synaptic noise, or architectural bottlenecks), discretization of the kind demonstrated here could contribute to categorical perception.

8.2 Working Memory Capacity

Miller's "7 ± 2" limit on working memory items [Miller, 1956] might reflect capacity constraints analogous to those in our simulation. If working memory operates through a noisy channel, the number of distinguishable items would be constrained by an information-capacity bound rather than arbitrary neural architecture.

8.3 Symbol Emergence in Language

Discrete tokens are fundamental to language. Our simulation illustrates one mechanism by which discretization can emerge without explicit symbolic machinery—through the interaction of compression and noise. This might inform theories of language evolution, though many other factors are surely involved.

8.4 Digital Information in Biology

Life uses digital information (genetic sequences) despite operating in an analog chemical medium. Noise and capacity constraints may have played a role in the emergence of discrete coding in early self-replicating systems [Walker and Davies, 2013, Deacon, 2011], though this remains highly speculative.

8.5 Relation to Thermodynamics

One might ask how this relates to Landauer's principle, which bounds the energy cost of bit erasure [Landauer, 1961, Bennett, 1982]. Our simulation has no explicit energy model, so we cannot make direct thermodynamic claims.

However, a thermodynamic extension is natural: treat the reconstruction loss as an energy-like potential $U(\mathbf{c})$ and the channel noise variance as proportional to temperature ($\sigma^2 \propto k_B T$). The system then minimizes an effective free energy $F = \langle U \rangle - TS$, where S is the entropy of the code distribution. Discretization would correspond to a "crystallization" into low-energy, high-entropy basins when temperature (noise) is high relative to channel capacity.

In this picture, the stochastic resonance phenomenon becomes standard barrier-crossing dynamics: noise provides the thermal energy to escape local minima and explore the potential landscape. The "information" emerges when the system settles into a basin—the uncertainty collapses, and the basin identity carries $\log_2 n$ bits.

We have not developed this thermodynamic model here; doing so properly would require specifying the potential landscape, computing Kramers rates for basin transitions, and tracking entropy production along trajectories. This is a natural direction for future work, connecting to stochastic thermodynamics and the physics of biological information processing. For now,

we simply note that the geometry demonstrated here—capacity-limited discretization, SR-like basin escape—is precisely what a thermodynamic model would need to explain.

9 Relation to Prior Work

The results connect to several established frameworks:

Rate–distortion theory and vector quantization [Cover and Thomas, 2006]: Our encoder-decoder system minimizes distortion subject to a rate constraint (channel capacity). At low rates, the optimal solution involves quantization—precisely the discrete codes we observe.

Information bottleneck [Tishby et al., 2000]: The bottleneck principle frames representation learning as extracting minimal sufficient statistics. Our simulation provides a concrete visualization of what happens geometrically when the bottleneck is tight.

What's new here: The explicit topology-change picture (ring $\rightarrow$ line/clusters/ring), the use of PR and ARI to track geometry and semantics, and the SR-like dimensionality story.

10 Limitations

This simulation is deliberately minimal:

- **Static data**: We use a fixed manifold, not dynamical or temporal structure
- **Gaussian noise**: Real channels have more complex noise statistics
- **Feedforward architecture**: Biological systems have recurrence and feedback
- **Single manifold**: The input is a simple ring; real data has more complex topology
- **No explicit physics**: There is no energy function or thermodynamic model

These limitations define the scope. Extending to time-series, recurrent architectures, and explicit energy models is future work.

11 Reproducibility

The simulation is implemented in Python using PyTorch. Code and figures are available at `https://github.com/todd866/code-formation`. Key parameters:

- Input: 2000 points, 512 dimensions, ring manifold (S^1)
- Encoder/Decoder: 3-layer MLP, 256 hidden units, ReLU activation
- Training: Adam optimizer, 150 epochs, MSE loss
- Critical experiment: $k = 2$, $\sigma = 0.5$

Random seeds are fixed for reproducibility. Running `python code_formation.py` generates all figures.

12 Conclusion

We have presented a minimal simulation demonstrating the emergence of discrete codes from continuous substrates when information passes through a noisy, capacity-limited channel. The key finding is that **both** a dimensional bottleneck **and** channel noise are required—neither alone suffices.

The mechanism is geometrically intuitive: noise creates uncertainty regions around each code point, and limited channel capacity means only a finite number of well-separated codes can coexist. The network discovers this constraint through gradient descent, producing discrete clusters with safety margins.

This illustrates one concrete mechanism by which symbol-like discrete codes can emerge from continuous substrates under capacity and noise constraints. Whether analogous mechanisms operate in biological systems—in perception, memory, language, or molecular information—remains an open and fascinating question.

References

Charles H Bennett. The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12):905–940, 1982.

John H Conway and Neil JA Sloane. *Sphere packings, lattices and groups*, volume 290. Springer Science & Business Media, 1999.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

Terrence W Deacon. Incomplete nature: How mind emerged from matter. *WW Norton & Company*, 2011.

Luca Gammaitoni, Peter Hänggi, Peter Jung, and Fabio Marchesoni. Stochastic resonance. *Reviews of Modern Physics*, 70(1):223, 1998.

Stevan Harnad. Categorical perception: The groundwork of cognition. *Cambridge University Press*, 1987.

Rolf Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81, 1956.

Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Sara Imari Walker and Paul CW Davies. The algorithmic origins of life. *Journal of the Royal Society Interface*, 10(79):20120869, 2013.