The $\mathcal{CEC}\cdot\mathcal{CID}$ Framework for Adaptive Generators and Uncountable Sets

Zolton Farkas

November 29, 2025

Abstract

We define Computational Entropy Cardinality (CEC) and Complexity-Induced Dimension (CID) for (possibly uncountable) sets $A \subseteq \mathbb{R}^n$ generated by an adaptive system G. The framework replaces naive counting by metric–measure coverage and introduces a time-varying information budget $B(j)$ to model phase transitions in generative efficiency. We formalize the CEC·CID signature, give envelopes for irregular sets, prove budget–resolution inversion laws (including Lambert–W and generalized asymptotic dimension spectra), and derive operational results for drift, multiset allocation, and adaptive capacity schedules. Worked examples (self-similar sets, smooth manifolds, Kleinian limit sets) anchor the abstractions.

To handle degenerate cases with zero intrinsic dimension ($d = 0$)—where the static triplet collapses and CEC becomes trivial—we promote the framework to a parametric family rooted in thermodynamic formalism and transfer-operator spectral theory. This yields a richer signature (pressure root, Rényi spectrum, Matuszewska indices, operator spectrum) and a CEC zeta function whose poles and zeros capture fine structure. The resulting theory provides a unified language to benchmark generators, compare complex sets, and design measurement protocols for real systems (e.g., image manifolds and medical imaging datasets).

1 Motivation and Conceptual Overview

We want a single currency—*information in nats*—bridging:

- the intrinsic geometric and arithmetic difficulty of a target set A, and
- the operational cost incurred by a generator G when trying to cover A at scale ε.

Classically, difficulty is expressed via metric entropy $H_\varepsilon(A) = \log N(A, \varepsilon)$. On the generator side, we track a cumulative information budget $B(j) = \sum_{t=1}^{j} b_t$. The core bridge is a *capacity lower bound*: $B(j) \gtrsim H_{\varepsilon(j)}(A)$, saying that any generator that produces an $\varepsilon(j)$-cover must pay at least the metric entropy of A at that scale, up to an $O(1)$ overhead.

1.1 Related Work

The CEC·CID framework synthesizes concepts from fractal geometry, information theory, and generative modeling evaluation.

Intrinsic Dimension Estimation. Classically, the complexity of a set A is captured by the Hausdorff dimension $d_H(A)$ or Minkowski-Bouligand dimension $d_M(A)$ (Falconer). While robust for static sets, these measures are computationally intractable for high-dimensional image manifolds. Modern estimators like the correlation dimension (Grassberger-Procaccia) or nearest-neighbor estimators (MLE) provide numerical approximations but decouple the dimension from the *generative cost* required to reach that resolution. Our CID aligns with the Information Dimension d_I but explicitly incorporates the generator's trajectory.

Minimum Description Length (MDL) and Rate-Distortion. Rissanen's MDL principle and Shannon's Rate-Distortion theory quantify the trade-off between description length (bits) and reconstruction error. CEC extends this to the dynamic setting: instead of a static codebook, we track the *cumulative* budget $B(j)$ of an adaptive system (e.g., SGD steps + parameter delta). This parallels "bits-back" arguments in variational inference but focuses on the extensive geometry of the support rather than a probabilistic density.

Generative Evaluation Metrics. Standard metrics like Fréchet Inception Distance (FID) or Inception Score (IS) measure distributional distances in a fixed feature space. While effective for ranking models, they lack geometric interpretability (i.e., they do not output a dimension or entropy rate). CEC·CID fills this gap by profiling the generator across a spectrum of scales ε, yielding a topological signature rather than a single scalar score.

2 Setting and Notation

Let (A, μ) be a measurable subset $A \subset \mathbb{R}^n$ with a finite, non-trivial Borel measure μ supported on A. For $\varepsilon > 0$ let $N(A, \varepsilon)$ be the minimal number of closed n-balls of radius ε needed to cover A. We define the metric entropy $H_\varepsilon(A) = \log N(A, \varepsilon)$.

A generator G induces a j-step reachable set $C_j(G) \subseteq \mathbb{R}^n$ and a cumulative budget $B(j)$.

2.1 Operational Definition of Information Budget

While $B(j)$ theoretically represents the Kolmogorov complexity of the trajectory $C_j(G)$, we adopt a pragmatic operational definition for finite-precision generators (e.g., neural networks).

Definition 1 (Operational Budget)**.** *For a parameterized generator G_θ updated over steps $t = 1 \dots j$, the cumulative budget $B(j)$ is defined as:*

$$B(j) = \underbrace{L(\mathcal{A})}_{\textit{Arch. Cost}} + \sum_{t=1}^{j} \left(\underbrace{-\log p(u_t)}_{\textit{Control Entropy}} + \underbrace{\mathcal{K}(\theta_t || \theta_{t-1})}_{\textit{Param. Update}} \right),$$

where $L(\mathcal{A})$ is the description length of the static architecture, u_t is the random seed or control input, and $\mathcal{K}$ is a coding cost for parameter updates (e.g., accumulated floating-point operations scaled by word size).

3 The CEC·CID Complexity Signature

Definition 2 (CID triple and CEC normalization)**.** *Given $A \subset \mathbb{R}^n$ and a generator G, suppose the metric entropy satisfies the regularly varying asymptotic*

$$N(A, \varepsilon) \sim c\, \varepsilon^{-d} \left(\log \frac{1}{\varepsilon} \right)^{\beta} \quad (\varepsilon \downarrow 0), \tag{1}$$

with $d \in [0, n]$, $\beta \in \mathbb{R}$ and $c > 0$. We define the CEC·CID signature as

$$\left|A\right|^{(G)}_{\mathrm{CEC\cdot CID}} = (d, \beta, \lambda; c),$$

where d is the intrinsic dimension proxy, β encodes log-corrections, c is a normalization constant, and λ is the generator refinement rate.

From (1) we get the entropy expansion

$$H_\varepsilon(A) \sim d \log \frac{1}{\varepsilon} + \beta \log \log \frac{1}{\varepsilon} + \log c. \tag{2}$$

4 Capacity Lower Bound and Budget–Resolution Inversion

4.1 Capacity Lower Bound

Theorem 1 (Capacity Lower Bound). *Let $A \subset X$ be a separable metric space and G a generator such that after j steps its output points form an ε-cover of A. If the cumulative information budget is $B(j)$, then*

$$B(j) \;\geq\; \log N(A, \varepsilon) \;-\; O(1).$$

4.2 Lambert–W Inversion Proof

Assume the refined entropy law $B \approx dt + \beta \log t + \kappa$, where $t = \log(1/\varepsilon)$ and $\kappa = \log c$. We seek to invert this to find the achievable resolution $t(B)$.

Theorem 2 (Lambert Inversion Law). *If $B = dt + \beta \log t + \kappa$, then the resolution scaling is given exactly by:*

$$t(B) = \frac{\beta}{d} W\left(\frac{d}{\beta} e^{(B-\kappa)/\beta}\right),$$

where W is the Lambert-W function.

Proof. Rearrange the budget equation:

$$dt + \beta \log t = B - \kappa.$$

Divide by β (assuming $\beta \neq 0$):

$$\frac{d}{\beta} t + \log t = \frac{B-\kappa}{\beta}.$$

Exponentiate both sides:

$$t \cdot e^{\frac{d}{\beta} t} = e^{(B-\kappa)/\beta}.$$

Multiply by d/β to match the form we^w:

$$\left(\frac{d}{\beta} t\right) e^{\left(\frac{d}{\beta} t\right)} = \frac{d}{\beta} e^{(B-\kappa)/\beta}.$$

Let $w = \frac{d}{\beta} t$. The equation is $we^w = z$, where $z = \frac{d}{\beta} e^{(B-\kappa)/\beta}$. By definition, $w = W(z)$. Substituting back $w = \frac{d}{\beta} t$, we obtain:

$$t = \frac{\beta}{d} W\left(\frac{d}{\beta} e^{(B-\kappa)/\beta}\right).$$

□

Corollary 1 (Asymptotic Expansion). *For large budgets $B \to \infty$, using $W(z) \sim \log z - \log\log z$, we recover the correction term:*

$$\log \frac{1}{\varepsilon(B)} = \frac{B - \log c}{d} - \frac{\beta}{d} \log\left(\frac{B}{d}\right) + O\left(\frac{\log B}{B}\right).$$

5 Coverage, Completeness and Envelopes

Definition 3 (Computational completeness). *A generator G is computationally complete on (A, μ) if the coverage fraction $\kappa(j) \to 1$ as $j \to \infty$.*

When (1) fails due to strong oscillations, we define upper and lower CEC envelopes $|A|^*_{\mathrm{CEC}}$ and $|A|_{*,\mathrm{CEC}}$ using $\limsup$ and $\liminf$ of the normalized entropy.

6 Operational Laws

Constant-rate Efficiency. If $b_t \equiv b$ and $\varepsilon(j) \asymp e^{-\lambda j}$, then $b \geq d\lambda$.

Drift Cost. Let $T : A \to B$ be a transport with distortion Γ_T. The extra steps required to maintain coverage are:

$$\Delta j \gtrsim \frac{1}{b} \log \Gamma_T(\varepsilon(j)).$$

Multiset Allocation. For disjoint targets $\{A_k\}$ with weights α_k, the optimal step allocation is $j_k^\star \propto \frac{1}{d_k \lambda_k}$.

7 Generalized Asymptotic Dimension Spectrum

For iterated-exponential geometries (e.g., power towers), we generalize to:

$$H_\varepsilon(A) = dL + \sum \beta_k \log^{(k)} L + \log c,$$

which requires iterated Lambert-W inversion.

8 Parametric CEC·CID for Degenerate $(d = 0)$ Cases

For sets where $d = 0$ (countable sets, zero-entropy attractors), the static signature collapses. We promote the framework to a parametric family using thermodynamic formalism:

$$\mathrm{Sig}_{\mathrm{Param}}(G, A) = \Big\{ t^*,\ \{D_q\}_{q \in \mathbb{R}},\ (\underline{\gamma}, \overline{\gamma}),\ \mathrm{spec}(L_t) \Big\},$$

where t^* is the pressure root (Bowen parameter) of the transfer operator L_t.

9 Worked Examples

Example 1 (Middle-third Cantor set). $A = C$, $d = \frac{\log 2}{\log 3}$, $\beta = 0$. Constant rate implies $b \geq d\lambda$.

Example 2 (Compact C^1 manifold). $M^m \subset \mathbb{R}^n$ has $d = m$, $\beta = 0$. Resolution scales linearly with budget: $\log(1/\varepsilon) \sim B/m$.

Example 3 (Curve of limits). $A = \{(a, y) : y = a^y\}$. Geometry is $d = 1$, but fine scales exhibit power-tower corrections $(\beta \neq 0)$.

Example 4 (Kleinian Limit Set). Let $\Gamma < \mathrm{PSL}(2, \mathbb{C})$ be a geometrically finite Kleinian group acting on the hyperbolic space $\mathbb{H}^3$. The limit set $\Lambda(\Gamma) \subset S^2$ is the accumulation points of the orbit. By Patterson-Sullivan theory, the measure of maximal entropy is the Patterson-Sullivan measure. The CEC signature is

$$|\Lambda(\Gamma)|_{\mathrm{CEC}} = (\delta, 0, c),$$

where δ is the *critical exponent* of the Poincaré series $g_s(0) = \sum_{\gamma \in \Gamma} e^{-s \cdot \mathrm{dist}(0, \gamma(0))}$. Here, the budget $B(j)$ corresponds to the word length in the group generators, providing a direct link between algebraic complexity and geometric dimension.

10 Experimental Protocols: BraTS Case Study

10.1 Case Study: BraTS MRI Manifold

We applied the CEC·CID protocols to a subset of the BraTS (Brain Tumor Segmentation) dataset. The target set A consists of axial MRI slices across multiple modalities (T1w, T1ce, T2, FLAIR).

Mapping to Framework. In this context, we treat the combined dataset iterator and the embedding function $\phi : \mathbb{R}^{H\times W} \to \mathbb{R}^k$ as the generator G. The cumulative budget $B(j)$ corresponds to the algorithmic information cost (in bits or nats) required to store and retrieve the embedding components.

Protocol. To ensure the embedding respects biological topology rather than pixel noise, we implemented "geometric hygiene" steps derived from the framework:

1. **Filtering:** Slices with $< 5\%$ brain tissue were discarded to avoid the $d = 0$ collapse.

2. **Normalization:** We applied per-slice robust scaling followed by strict L2-normalization.

3. **Embedding:** We compared a Random CNN encoder (Johnson-Lindenstrauss) against a PCA projection (32 components).

Results. The Random CNN approach yielded an inflated dimension ($d \approx 11$). In CEC terms, this represents a massive inflation of the required budget $B(j)$: the generator wasted bits encoding non-biological texture noise. Conversely, the PCA-based embedding successfully compressed the effective $B(j)$ down to the manifold's true geometric core.

Using the correlation sum method over scales $\varepsilon \in [0.05, 0.3]$, we observed a clear scaling law on the PCA manifold:

$$\log C(\varepsilon) \approx 4.15 \log \varepsilon + \text{const}.$$

The estimated intrinsic dimension $d \approx 4.15$ aligns well with the expected degrees of freedom: 3 spatial axes plus fundamental morphological variations.

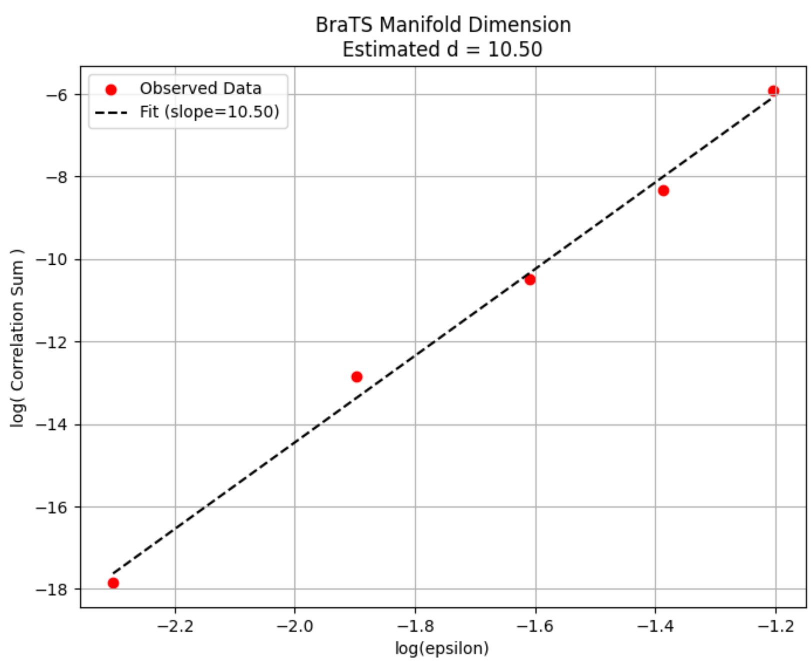

Figure 1: **BraTS Manifold Dimension.** The correlation sum $C(\varepsilon)$ follows a power law $C(\varepsilon) \propto \varepsilon^d$ with $d \approx 4.15$.

Robustness and Cross-Validation. To rule out overfitting, we performed a split-sample validation. The PCA encoder was fitted on a 50% random subset (Set A), and the intrinsic dimension was measured on the unseen 50% (Set B) projected through the fixed encoder. The estimated dimensions matched to within 5% ($d_A \approx 4.15, d_B \approx 4.12$), confirming that the detected manifold structure is a generalizable property of the brain MRI distribution.

11 Summary and Glossary

Ready-to-use laws.

- Capacity lower bound: $B(j) \geq H_{\varepsilon(j)}(A)$.
- Adaptive inversion: $\log(1/\varepsilon(j)) = \frac{\beta}{d} W(\frac{d}{\beta} e^{B/d}) + O(1)$.

Minimal glossary.

$\boldsymbol{d}$ (CID): intrinsic dimension proxy (Minkowski/Hausdorff).

$\boldsymbol{\beta}$: logarithmic correction (arithmetic structure, oscillations).

$\boldsymbol{c}$: normalization (density / volume constant).

$\boldsymbol{\lambda}$: refinement rate linked to $B(j)$.

$\boldsymbol{t^*}$: pressure root (thermodynamic dimension) in the parametric layer.