Excess Description Length: An Information Measure for Generalizable Structure Learned from Finite Data

Elizabeth Donoway
Department of Physics
University of California, Berkeley
Berkeley, CA, USA
donoway@berkeley.edu

***Abstract*—Learning algorithms transfer predictive structure from data into model parameters, yet existing information-theoretic quantities primarily characterize either dataset complexity or asymptotic learning costs. We introduce *excess description length* (EDL), a compute-indexed measure of the generalizable information extracted by a learning algorithm from finite data. EDL admits an operational interpretation via prequential coding: it is the excess codelength incurred on first exposure to data relative to a predictor whose expected log-loss equals the final model's *population* loss, thereby excluding memorization effects on the training set. We establish that EDL satisfies fundamental properties of an information measure—non-negativity, additivity under product structure, monotonicity in compute, and a processing bound on extractable information—while explicitly decoupling the roles of data (information source) and computation (extraction mechanism). A finite-data saturation bound formalizes that compute cannot extract more generalizable information than exists in the data. We show that EDL penalizes overfitting through the generalization gap, distinguishing genuine learning from memorization. These results establish EDL as a rigorous foundation for quantifying information transfer in learning systems operating on finite datasets.**

***Index Terms*—Excess description length, minimum description length, prequential coding, generalization, information measures**

I. Introduction

Information theory provides tools for quantifying uncertainty in data and complexity in models, yet learning algorithms occupy an intermediate role: they transform finite datasets into predictors that generalize beyond observed samples. Classical quantities such as entropy, mutual information, and minimum description length (MDL) characterize properties of data or hypothesis classes [1], [2], [4], while recent work explores dataset structure under computational constraints [8], [9]. However, these frameworks do not directly measure how much *generalizable* predictive information a learning algorithm extracts from finite data as a function of computation.

Modern learning commonly uses computation exceeding dataset size, revisiting the same finite data over many epochs. In this regime, naive prequential codes that charge loss for every repeated example conflate computation with new information. We seek a quantity that:

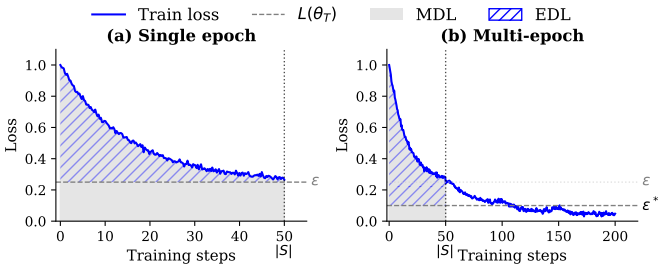

Fig. 1. **EDL measures generalizable information extracted from finite data.** The data S (size $|S|$) are fixed. (a) After one epoch, EDL (blue hatched) reflects structure learned from S so far. (b) Multi-epoch training on the same data improves generalization ($L(\theta_T) = \varepsilon^* < \varepsilon$), yielding larger EDL. Total area under the first-epoch curve (up to $|S|$) is MDL (gray); EDL is the portion above final population loss $L(\theta_T)$. Additional training extracts more of the learnable structure in S without adding new information.

1) Measures generalizable (population-level) information, not training fit;
2) Explicitly indexes computation, allowing multi-epoch training;
3) Charges information only for first exposure to each datum; and
4) Satisfies axiomatic properties expected of information measures.

We define *excess description length* (EDL), a compute-indexed quantity measuring how much population-predictive structure is transferred from data into model parameters. EDL equals the prequential codelength incurred during a single pass through the data, minus the codelength that would be required using the final trained predictor. This "first-exposure area above final loss" (Fig. 1) captures information absorbed into the model during learning.

Our main contributions are:

- A formal definition of EDL that decouples data from computation (Section II);
- Proofs that EDL satisfies non-negativity, additivity, processing bounds, and compute monotonicity (Section III);

- A saturation theorem showing computation cannot extract more information than exists in finite data (Section IV);
- Analysis of how EDL distinguishes generalization from memorization (Section V);
- Connections to regret, MDL, surplus description length (SDL), and concurrent work on epiplexity (Section VI).

An extended companion paper [10] applies the EDL framework to analyze capability emergence in language models.

II. Setup and Definitions

Let $(X, Y) \sim \mathcal{D}$ be drawn from a data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Consider a conditional model family $\{p_\theta(y|x) : \theta \in \Theta\}$. We use log-loss (in bits): $\ell(\theta; x, y) \triangleq -\log_2 p_\theta(y|x)$. The *population loss* is

$$L(\theta) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\theta; x, y)]. \tag{1}$$

Let $S = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$ be a training set. A (possibly randomized) learning algorithm $\mathcal{A}$ maps an initialization θ_0 and data S to parameters $\theta_T = \mathcal{A}(\theta_0, S; T)$ after compute budget T (measured in optimization steps, epochs, or FLOPs).

A. First-Exposure Prequential Codelength

To define a codelength that does not scale with repeated epochs, we use the *first-exposure sequence* from a single pass through a random permutation of the data.

Let π be a uniform random permutation of $[n]$, and let S_π denote the ordered sequence $((x_{\pi(1)}, y_{\pi(1)}), \ldots, (x_{\pi(n)}, y_{\pi(n)}))$. Let θ_i^π denote the parameters after processing the first i examples once.

Definition 1 (First-exposure prequential codelength)**.**

$$\mathrm{MDL}_1(S; \mathcal{A}) \triangleq \mathbb{E}_\pi \left[\sum_{i=1}^n \ell(\theta_{i-1}^\pi; x_{\pi(i)}, y_{\pi(i)}) \right]. \tag{2}$$

This is the expected codelength to encode the training labels using the evolving model during first exposure, averaging over presentation orders.

Remark 1 (Implementation)**.** *For SGD-style methods, θ_i^π is the state after i minibatches in the first epoch. More generally, θ_i^π can be defined by running $\mathcal{A}$ on prefixes or via a specified online variant. The results below depend only on the induced sequence $\{\theta_i^\pi\}_{i=0}^n$.*

B. Excess Description Length

Definition 2 (Compute-indexed excess description length)**.** *For compute budget T, define the (population) EDL:*

$$\mathrm{EDL}_T(\mathcal{A}, n) \triangleq \mathbb{E}_{S\sim\mathcal{D}^n} \left[\mathrm{MDL}_1(S; \mathcal{A}) - n \cdot L(\theta_T) \right], \tag{3}$$

where $\theta_T = \mathcal{A}(\theta_0, S; T)$.

This definition formalizes the operational estimator "area of the first-epoch prequential loss above the final population loss," while allowing θ_T to come from multi-epoch training ($T \gg n$).

Remark 2 (Sign of EDL)**.** *Unlike entropy, EDL can be negative when the learning algorithm degrades population-level predictions (see Corollary 4). Positive EDL indicates successful extraction of generalizable structure; negative EDL indicates overfitting or other pathologies.*

Remark 3 (Operational estimation)**.** *In practice, EDL_T is estimated by: (1) computing cumulative loss during the first epoch of training, (2) estimating $L(\theta_T)$ on held-out data after training completes, and (3) taking the difference.*

III. Properties as an Information Measure

We establish that EDL_T satisfies properties expected of an information measure, paralleling Shannon's axioms while explicitly parameterizing computation.

A. Prequential Interpretation

Theorem 1 (Expected area form)**.** *Assume $S \sim \mathcal{D}^n$ and π is uniform over permutations. Then*

$$\mathrm{EDL}_T(\mathcal{A}, n) = \sum_{i=1}^n \mathbb{E}\left[L(\theta_{i-1}^\pi) - L(\theta_T) \right]. \tag{4}$$

Proof. By the tower property, conditioning on θ_{i-1}^π:

$$\mathbb{E}[\ell(\theta_{i-1}^\pi; x_{\pi(i)}, y_{\pi(i)}) \mid \theta_{i-1}^\pi] = L(\theta_{i-1}^\pi),$$

since $(x_{\pi(i)}, y_{\pi(i)}) \sim \mathcal{D}$ is independent of θ_{i-1}^π (which depends only on earlier examples). Summing over i and subtracting $n \cdot L(\theta_T)$ gives the result. □

This expresses EDL as the cumulative excess population loss along the first-exposure trajectory relative to the final model.

Remark 4 (Algorithm-dependence as a feature)**.** *The expected area form makes explicit that EDL depends not only on the final model, but also on the learning trajectory. Two algorithms that reach the same θ_T have different EDL if one incurs larger population loss during first exposure. This is by design: EDL measures the efficiency with which an algorithm extracts generalizable structure from data; it is not an algorithm-independent notion of final information content.*

An algorithm that learns slowly incurs high regret and thus high EDL, reflecting inefficient compression of the data stream. If one desires an intrinsic quantity independent of learning dynamics, the appropriate object is $\sup_{\mathcal{A}} \mathrm{EDL}_T(\mathcal{A}, n)$, which optimizes over algorithms such that optimal learners dominate.

B. Non-Negativity

Definition 3 (Population-monotone algorithm)**.** *Algorithm $\mathcal{A}$ is* population-monotone *if along the induced trajectory, $\mathbb{E}[L(\theta_{t+1})] \le \mathbb{E}[L(\theta_t)]$ for all t.*

This condition holds for gradient descent on convex losses, SGD with sufficiently small learning rate, and algorithms with early stopping that prevents overfitting.

Theorem 2 (Non-negativity)**.** *If $\mathcal{A}$ is population-monotone and T is at least the length of the first-exposure trajectory, then $\mathrm{EDL}_T(\mathcal{A}, n) \ge 0$.*

Proof. Population-monotonicity implies $\mathbb{E}[L(\theta_T)] \leq \mathbb{E}[L(\theta^\pi_{i-1})]$ for each $i \leq n$. From Theorem 1, each term $\mathbb{E}[L(\theta^\pi_{i-1}) - L(\theta_T)] \geq 0$, so the sum is non-negative. □

C. Monotonicity in Compute

Theorem 3 (Compute monotonicity)**.** *If $T_2 \geq T_1 \geq n$ and $\mathbb{E}[L(\theta_{T_2})] \leq \mathbb{E}[L(\theta_{T_1})]$, then*

$$\mathrm{EDL}_{T_2}(\mathcal{A}, n) \geq \mathrm{EDL}_{T_1}(\mathcal{A}, n),$$

with equality iff $\mathbb{E}[L(\theta_{T_2})] = \mathbb{E}[L(\theta_{T_1})]$.

Proof. Since MDL_1 depends only on the first-exposure trajectory (independent of T for $T \geq n$):

$$\mathrm{EDL}_{T_2} - \mathrm{EDL}_{T_1} = n(\mathbb{E}[L(\theta_{T_1})] - \mathbb{E}[L(\theta_{T_2})]) \geq 0.$$

□

Corollary 4 (Overfitting decreases EDL)**.** *For $T \geq n$, if additional compute increases population loss (overfitting), EDL decreases. Thus EDL_T is maximized at the early stopping point $T^* = \arg\min_T \mathbb{E}[L(\theta_T)]$.*

Remark 5 (Negative EDL)**.** *When overfitting is severe—specifically, when $n \cdot L(\theta_T) > \mathrm{MDL}_1(S; \mathcal{A})$—EDL becomes negative. Though not a valid codelength, negative EDL indicates that training has degraded the model's population-level predictions relative to its earlier state, indicating that the algorithm has "unlearned" generalizable structure.*

D. Additivity

We first show that EDL decomposes into per-example contributions expressible as KL divergence reductions.

Theorem 5 (KL decomposition)**.** *EDL decomposes as:*

$$\mathrm{EDL}_T(\mathcal{A}, n) = \sum_{i=1}^{n} \mathbb{E}\left[\Delta_{\mathrm{KL}}(\theta^\pi_{i-1}, \theta_T)\right], \tag{5}$$

where $\Delta_{\mathrm{KL}}(\theta, \theta') = \mathrm{D}_{\mathrm{KL}}(p_\mathcal{D} \| p_\theta) - \mathrm{D}_{\mathrm{KL}}(p_\mathcal{D} \| p_{\theta'})$ is the KL improvement, with $p_\mathcal{D}(y|x)$ the true conditional.

Proof. Using $L(\theta) = H_\mathcal{D}(Y|X) + \mathrm{D}_{\mathrm{KL}}(p_\mathcal{D} \| p_\theta)$, we have:

$$L(\theta^\pi_{i-1}) - L(\theta_T) = \mathrm{D}_{\mathrm{KL}}(p_\mathcal{D} \| p_{\theta^\pi_{i-1}}) - \mathrm{D}_{\mathrm{KL}}(p_\mathcal{D} \| p_{\theta_T}).$$

The result follows from Theorem 1. □

Theorem 6 (Additivity for independent data)**.** *Let $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2$ over $(\mathcal{X}_1, \mathcal{Y}_1) \times (\mathcal{X}_2, \mathcal{Y}_2)$, and let $S_1 \sim \mathcal{D}_1^{n_1}$ and $S_2 \sim \mathcal{D}_2^{n_2}$ be independent. Suppose the model factors as $p_\theta(y_1, y_2|x_1, x_2) = p_{\theta^{(1)}}(y_1|x_1) \cdot p_{\theta^{(2)}}(y_2|x_2)$ and $\mathcal{A}$ respects this factorization. Then:*

$$\mathrm{EDL}_T(\mathcal{A}, S_1 \cup S_2) = \mathrm{EDL}_{T_1}(\mathcal{A}_1, n_1) + \mathrm{EDL}_{T_2}(\mathcal{A}_2, n_2). \tag{6}$$

Proof. Under the factorization, losses on $\mathcal{D}_1$ examples depend only on $\theta^{(1)}$ and vice versa. Thus:

$$\mathrm{MDL}_1(S_1 \cup S_2; \mathcal{A}) = \mathrm{MDL}_1(S_1; \mathcal{A}_1) + \mathrm{MDL}_1(S_2; \mathcal{A}_2),$$
$$L(\theta_T) = L_1(\theta_T^{(1)}) + L_2(\theta_T^{(2)}).$$

Subtracting gives the claimed decomposition. □

E. Decomposition for Sequential Learning

We now consider settings where data arrive in stages, as in curriculum learning or continual learning. Unlike the additivity result (Theorem 6), which applies to independent data with factored models, the following decomposition applies to data from the same distribution processed sequentially.

Definition 4 (Warm-start learning)**.** *Given datasets S_1 and S_2 and compute budgets T_1, T_2, a* warm-start *procedure:*

1) *Trains on S_1 for T_1 steps, yielding θ_{T_1};*
2) *Initializes from θ_{T_1} and trains on S_2 for T_2 steps, yielding θ_T.*

We write $\mathcal{A}_{|S_1 \to S_2}$ for such a procedure and $\mathrm{EDL}_{T_2|\theta_{T_1}}(\mathcal{A}, S_2)$ for the EDL of the second stage.

Theorem 7 (Sequential decomposition)**.** *Let $S_1 \sim \mathcal{D}^{n_1}$ and $S_2 \sim \mathcal{D}^{n_2}$ be independent samples from the same distribution $\mathcal{D}$. For a warm-start procedure $\mathcal{A}_{|S_1 \to S_2}$:*

$$\begin{aligned}\mathrm{EDL}_T(\mathcal{A}, S_1 \cup S_2) = \mathrm{EDL}_{T_1}(\mathcal{A}, n_1) + \mathrm{EDL}_{T_2|\theta_{T_1}}(\mathcal{A}, n_2) \\ + n_1 \cdot \mathbb{E}[L(\theta_{T_1}) - L(\theta_T)],\end{aligned} \tag{7}$$

where the final term is the transfer benefit*: the improvement in population loss from training on S_2 after S_1, scaled by total sample size.*

Proof. The total first-exposure codelength decomposes as:

$$\mathrm{MDL}_1(S_1 \cup S_2; \mathcal{A}) = \mathrm{MDL}_1(S_1; \mathcal{A}) + \mathrm{MDL}_1(S_2; \mathcal{A}_{|\theta_{T_1}}),$$

where $\mathcal{A}_{|\theta_{T_1}}$ denotes the algorithm initialized at θ_{T_1}.

The total EDL is:

$$\begin{aligned}\mathrm{EDL}_T &= \mathrm{MDL}_1(S_1 \cup S_2; \mathcal{A}) - (n_1 + n_2)L(\theta_T) \\ &= [\mathrm{MDL}_1(S_1; \mathcal{A}) - n_1 L(\theta_{T_1})] \\ &\quad + \left[\mathrm{MDL}_1(S_2; \mathcal{A}_{|\theta_{T_1}}) - n_2 L(\theta_{T_1})\right] \\ &\quad + (n_1 + n_2)L(\theta_{T_1}) - (n_1 + n_2)L(\theta_T).\end{aligned}$$

The first bracket is $\mathrm{EDL}_{T_1}(\mathcal{A}, n_1)$. For the second bracket, note that $\mathrm{EDL}_{T_2|\theta_{T_1}}(\mathcal{A}, n_2)$ as standardly defined would subtract $n_2 L(\theta_T)$, not $n_2 L(\theta_{T_1})$. We have:

$$\begin{aligned}&\mathrm{MDL}_1(S_2; \mathcal{A}_{|\theta_{T_1}}) - n_2 L(\theta_{T_1}) \\ &= \left[\mathrm{MDL}_1(S_2; \mathcal{A}_{|\theta_{T_1}}) - n_2 L(\theta_T)\right] + n_2[L(\theta_T) - L(\theta_{T_1})] \\ &= \mathrm{EDL}_{T_2|\theta_{T_1}}(\mathcal{A}, n_2) + n_2[L(\theta_T) - L(\theta_{T_1})].\end{aligned}$$

Substituting back:

$$\begin{aligned}\mathrm{EDL}_T &= \mathrm{EDL}_{T_1}(S_1) + \mathrm{EDL}_{T_2|\theta_{T_1}}(S_2) \\ &\quad + n_2[L(\theta_T) - L(\theta_{T_1})] \\ &\quad + (n_1 + n_2)[L(\theta_{T_1}) - L(\theta_T)] \\ &= \mathrm{EDL}_{T_1}(S_1) + \mathrm{EDL}_{T_2|\theta_{T_1}}(S_2) \\ &\quad + (n_1 + n_2 - n_2)[L(\theta_{T_1}) - L(\theta_T)] \\ &= \mathrm{EDL}_{T_1}(S_1) + \mathrm{EDL}_{T_2|\theta_{T_1}}(S_2) \\ &\quad + n_1[L(\theta_{T_1}) - L(\theta_T)].\end{aligned}$$

□

Remark 6 (Transfer benefit)**.** *The term $n_1[L(\theta_{T_1}) - L(\theta_T)]$ quantifies how subsequent training on S_2 can retroactively increase the effective information extracted from S_1. When continued training reduces population loss, the original data is reinterpreted as having contributed more generalizable structure. This formalizes the intuition that curriculum strategies and staged training can enable more efficient structure extraction. Quantitative applications to pretraining and finetuning appear in [10].*

F. Mutual Information Bound

Lemma 8 (EDL bounded by mutual information)**.** *For any distribution $\mathcal{D}$, sample size n, and population-monotone algorithm $\mathcal{A}$ with* marginal-calibrated initialization *(i.e., $p_{\theta_0}(y|x) = p(y)$ so that $L(\theta_0) = H_{\mathcal{D}}(Y)$):*

$$\mathrm{EDL}_T(\mathcal{A}, n) \le n \cdot I_{\mathcal{D}}(X; Y). \tag{8}$$

Proof. From Theorem 1: $\mathrm{EDL}_T(\mathcal{A}, n) = \sum_{i=1}^{n} \mathbb{E}[L(\theta_{i-1}^{\pi}) - L(\theta_T)]$.

By marginal-calibrated initialization, $L(\theta_0) = H_{\mathcal{D}}(Y)$. Population-monotonicity ensures $L(\theta_{i-1}^{\pi}) \le L(\theta_0) = H_{\mathcal{D}}(Y)$ for all i. The minimum achievable final loss is $L^* = H_{\mathcal{D}}(Y|X)$.

Therefore, each term satisfies:

$$\mathbb{E}[L(\theta_{i-1}^{\pi}) - L(\theta_T)] \le H_{\mathcal{D}}(Y) - H_{\mathcal{D}}(Y|X) = I_{\mathcal{D}}(X; Y).$$

Summing over $i = 1, \ldots, n$ gives the result. The bound is tight when the final predictor achieves Bayes-optimal loss. □

Remark 7. *Marginal-calibrated initialization is standard practice and information-theoretically optimal when no data have been observed. Without this assumption, the bound becomes $\mathrm{EDL}_T \le n(L(\theta_0) - L^*)$.*

G. Processing Bound

We establish that data processing cannot increase the supremal extractable information.

Definition 5 (Representation-restricted algorithms)**.** *Let $f : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Z}$ be any measurable mapping from datasets to representations. Let $\mathcal{A}_f$ denote the class of learning procedures that access the data only through $f(S)$.*

Theorem 9 (Processing bound)**.** *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}' \times \mathcal{Y}'$ be a deterministic transformation applied elementwise, with pushforward distribution $\mathcal{D}' = f_*(\mathcal{D})$. Then:*

$$\sup_{\mathcal{A}'} \mathrm{EDL}_T(\mathcal{A}', f(S)) \le \sup_{\mathcal{A}} \mathrm{EDL}_T(\mathcal{A}, S). \tag{9}$$

Proof. Any algorithm $\mathcal{A}'$ on $f(S)$ can be simulated by $\tilde{\mathcal{A}}$ on S that first computes $f(S)$ internally. The simulator achieves identical MDL_1 and final predictor, hence identical EDL. Since $\tilde{\mathcal{A}}$ is valid on S:

$$\mathrm{EDL}_T(\mathcal{A}', f(S)) = \mathrm{EDL}_T(\tilde{\mathcal{A}}, S) \le \sup_{\mathcal{A}} \mathrm{EDL}_T(\mathcal{A}, S).$$

Taking the supremum over $\mathcal{A}'$ yields (9). □

Corollary 10 (Sufficient statistic invariance)**.** *If $f(x, y) = (T(x), y)$ where $T(x)$ is a sufficient statistic for Y given X, then:* $\sup_{\mathcal{A}'} \mathrm{EDL}_T(\mathcal{A}', f(S)) = \sup_{\mathcal{A}} \mathrm{EDL}_T(\mathcal{A}, S)$.

Remark 8 (Information lost in processing)**.** *The gap between the two sides of (9) quantifies predictive information destroyed by f. When f is lossy, no algorithm—regardless of compute—can recover the lost structure.*

IV. Finite-Data Saturation

A fundamental property of EDL is that computation cannot extract more generalizable information than exists in the data.

Theorem 11 (Saturation bound)**.** *For any population-monotone algorithm with marginal-calibrated initialization:*

$$\sup_{T} \mathrm{EDL}_T(\mathcal{A}, n) \le n \cdot I_{\mathcal{D}}(X; Y), \tag{10}$$

where $I_{\mathcal{D}}(X; Y) = H_{\mathcal{D}}(Y) - H_{\mathcal{D}}(Y|X)$ is the mutual information.

Proof. Immediate from Lemma 8. □

Corollary 12 (Decoupling data from compute)**.** *Let $T_1 = n$ (end of first epoch) and $T_2 > T_1$. Then:*

$$\mathrm{EDL}_{T_2} = \mathrm{EDL}_{T_1} + n \cdot \mathbb{E}[L(\theta_{T_1}) - L(\theta_{T_2})]. \tag{11}$$

Equation (11) shows that repeated optimization on finite data increases EDL *only* through further reductions in population loss. No resampled example is redundantly encoded as additional information. Multi-epoch training that improves generalization extracts more of the structure already present in S, without inventing spurious information.

V. Memorization vs. Generalization

EDL automatically distinguishes learning from memorization through the generalization gap.

Theorem 13 (Generalization gap penalty)**.** *Consider algorithms $\mathcal{A}_{\mathrm{gen}}$ achieving $L(\theta_T) \to L^*$ and $\mathcal{A}_{\mathrm{mem}}$ achieving zero training loss but $L(\theta_T) = L_0 > L^*$ (memorization without generalization). Assuming similar first-exposure trajectories:*

$$\mathrm{EDL}_T(\mathcal{A}_{\mathrm{gen}}) - \mathrm{EDL}_T(\mathcal{A}_{\mathrm{mem}}) = n(L_0 - L^*) > 0. \tag{12}$$

Proof. If both have similar initial behavior, $\mathrm{MDL}_1(S; \mathcal{A}_{\mathrm{gen}}) \approx \mathrm{MDL}_1(S; \mathcal{A}_{\mathrm{mem}})$. The EDL difference is: $n \cdot L(\theta_T^{\mathrm{mem}}) - n \cdot L(\theta_T^{\mathrm{gen}}) = n(L_0 - L^*) > 0$. □

Remark 9. *The assumption of similar first-exposure trajectories holds when both algorithms use the same architecture, initialization, and optimization during the first epoch, differing only in regularization or training duration.*

Corollary 14 (Random labels yield zero EDL)**.** *For data with random labels $Y \perp X$, $\mathrm{EDL}_T(\mathcal{A}, n) \to 0$ for any algorithm, regardless of training loss achieved.*

Proof. With $Y \perp X$, the optimal predictor is $p^*(y|x) = p(y)$, achieving $L^* = H(Y)$. No algorithm can improve on marginal

prediction for the population, so $L(\theta_T) \geq H(Y)$. A well-calibrated algorithm achieves $\mathrm{MDL}_1 \approx nH(Y)$, yielding $\mathrm{EDL}_T \approx 0$. □

This validates EDL as measuring learnable structure: random labels contain no generalizable information, and EDL correctly reports zero regardless of training effort.

VI. Connections to Prior Work

Table I compares EDL to related information measures.

MDL and prequential coding. The prequential approach to MDL [3], [5] uses cumulative predictive loss as description length. EDL extends this by subtracting the final model's population loss, isolating information *absorbed into parameters* from residual encoding cost.

Surplus description length. SDL [6] measures the asymptotic cost of learning an optimal predictor. Under consistency ($L(\theta_T) \to L^*$ as $n, T \to \infty$), EDL converges to SDL. However, EDL explicitly indexes computation and remains meaningful when $T \gg n$ (multi-epoch training), where SDL is not directly defined.

Regret. The first-exposure codelength relates to online learning regret [7]. Define regret relative to θ_T: $R_T(S_\pi) = \sum_{i=1}^{n} [\ell(\theta_{i-1}^{\pi}; z_{\pi(i)}) - \ell(\theta_T; z_{\pi(i)})]$. Then $\mathrm{EDL}_T = \mathbb{E}_{S,\pi}[R_T(S_\pi)] + n(\mathbb{E}[L_{\mathrm{train}}(\theta_T)] - L(\theta_T))$, connecting EDL to expected regret plus a generalization correction.

Epiplexity. Concurrent independent work [9] introduces "epiplexity" S_T as a compute-bounded notion of structural content in data. Given compute bound T, epiplexity measures the description length of the *optimal* model that can be trained *and* evaluated within T.

EDL differs in three respects: (i) it evaluates a *specified algorithm* rather than optimizing over all feasible programs to evaluate the *data*, (ii) it uses an explicit population-loss reference to exclude memorization, and (iii) it remains meaningful when compute exceeds dataset size ($T \gg n$) by fixing a first-exposure codelength.

In the one-pass i.i.d. regime, the prequential *proxy* for the epiplexity estimator takes an "area above final loss" form that is geometrically similar to single-epoch instantiations of EDL_T when the empirical risk coincides with the population risk. Our framework extends naturally to multi-epoch training: $\mathrm{EDL}_{T_{\mathrm{converged}}} > \mathrm{EDL}_{T_{\mathrm{single-epoch}}}$ when additional epochs improve generalization.

VII. Discussion

We have introduced excess description length as a rigorous measure of generalizable information extracted by learning algorithms from finite data. The central innovations are:

1) **Decoupling data from compute**: EDL counts information only for first exposure, while allowing arbitrary subsequent computation to extract structure.
2) **Population-level measurement**: By referencing population loss rather than training loss, EDL measures *generalizable* information, automatically penalizing overfitting.
3) **Axiomatic foundation**: Non-negativity, additivity, processing bounds, and saturation establish EDL as a proper information measure.

TABLE I
Comparison of information measures for learning

Property	H	K	SDL	S_T	**EDL**
Non-negative	✓	✓	✓	✓	✓
Additive	✓	$\pm O(1)$	?	?	✓
Processing bound	✓	✓	?	?	✓
Finite data	–	✓	×	✓*	✓
Multi-epoch	–	–	×	×	✓
Compute-indexed	×	×	×	✓	✓
Operationally computable	✓	×	Asymp.	Bounded	✓

H: Shannon entropy; K: Kolmogorov complexity; S_T: Epiplexity

*Epiplexity assumes data are sampled from an effectively infinite distribution with small generalization gap; sufficiently large compute bounds (exceeding the domain of the distribution) can violate this assumption.

The supremum $\sup_{\mathcal{A}} \lim_{T\to\infty} \mathrm{EDL}_T(\mathcal{A}, n)$ measures the maximum generalizable structure any algorithm can extract from n samples, in the limit of unlimited compute. This defines an intrinsic property of $(\mathcal{D}, n)$ analogous to channel capacity; characterizing this quantity remains open.

Practical implications. EDL provides a foundation for: (1) evaluating learning algorithms by information extraction efficiency [10], (2) understanding when multi-epoch training is beneficial versus when it overfits, and (3) quantifying the value of data augmentation and curriculum strategies.

Acknowledgment

E.D. thanks John Schulman, Jan Leike, Hailey Joren, Ethan Perez, Michael R. DeWeese, Fabien Roger, and Eric Easley for helpful discussions and feedback.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.
[2] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.
[3] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 629–636, 1984.
[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
[5] A. P. Dawid, "Statistical theory: The prequential approach," *J. Royal Statistical Society A*, vol. 147, no. 2, pp. 278–292, 1984.
[6] W. F. Whitney, M. J. Song, D. Brandfonbrener, J. Altosaar, and K. Cho, "Evaluating representations by the complexity of learning low-loss predictors," *arXiv:2009.07368*, 2021.
[7] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.
[8] Y. Xu, S. Zhao J. Song, R. Stewart, and S. Ermon, "A Theory of Usable Information under Computational Constraints," *International Conference on Learning Representations*, 2020.
[9] 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*, 2026.
[10] E. Donoway, H. Joren, F. Roger, J. Leike, "Excess description length of learning generalizable predictors," *arXiv:2601.04728*, 2026.

APPENDIX A
TRANSFER LEARNING DECOMPOSITION

The sequential decomposition (Theorem 7) enables a precise analysis of how pretraining affects fine-tuning information requirements. We state a rigorous version of the transfer learning result, making explicit the assumptions that govern when pretraining reduces fine-tuning EDL.

Corollary 15 (Transfer learning decomposition)**.** *Consider fine-tuning on $S_{\mathrm{ft}} \sim \mathcal{D}_{\mathrm{ft}}^{n_{\mathrm{ft}}}$ from either:*

- *Random initialization θ_0 (cold start), yielding first-exposure trajectory $(\theta_i^{\mathrm{cold}})_{i=0}^{n_{\mathrm{ft}}}$ and final model θ_T^{cold}; or*
- *Pretrained initialization $\theta_{T_{\mathrm{pre}}}$ (warm start), yielding trajectory $(\theta_i^{\mathrm{warm}})_{i=0}^{n_{\mathrm{ft}}}$ and final model θ_T^{warm}.*

Let $\mathrm{EDL}^{\mathrm{cold}} \triangleq \mathrm{EDL}_{T|\theta_0}(\mathcal{A}, S_{\mathrm{ft}})$ *and* $\mathrm{EDL}^{\mathrm{warm}} \triangleq \mathrm{EDL}_{T|\theta_{T_{\mathrm{pre}}}}(\mathcal{A}, S_{\mathrm{ft}})$ *denote the respective fine-tuning EDLs. Then:*

$$\mathrm{EDL}^{\mathrm{cold}} - \mathrm{EDL}^{\mathrm{warm}} = \underbrace{\sum_{i=1}^{n_{\mathrm{ft}}} \mathbb{E}\left[L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{cold}}) - L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{warm}})\right]}_{\textit{trajectory advantage } \Delta_{\mathrm{traj}}} + \underbrace{n_{\mathrm{ft}} \cdot \mathbb{E}\left[L_{\mathrm{ft}}(\theta_T^{\mathrm{warm}}) - L_{\mathrm{ft}}(\theta_T^{\mathrm{cold}})\right]}_{\textit{endpoint correction } \Delta_{\mathrm{end}}}. \tag{13}$$

The following special cases clarify when pretraining reduces fine-tuning EDL:

(i) ***Identical convergence.*** *If both procedures converge to the same final population loss, i.e.,* $\mathbb{E}[L_{\mathrm{ft}}(\theta_T^{\mathrm{warm}})] = \mathbb{E}[L_{\mathrm{ft}}(\theta_T^{\mathrm{cold}})]$*, then*

$$\mathrm{EDL}^{\mathrm{cold}} - \mathrm{EDL}^{\mathrm{warm}} = \Delta_{\mathrm{traj}} = \sum_{i=1}^{n_{\mathrm{ft}}} \mathbb{E}\left[L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{cold}}) - L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{warm}})\right]. \tag{14}$$

(ii) ***Trajectory dominance.*** *Under the additional assumption that the pretrained initialization yields uniformly lower population loss along the entire first-exposure trajectory (i.e.,* $L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{warm}}) \le L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{cold}})$ *almost surely for all* $i \le n_{\mathrm{ft}}$*) we have* $\Delta_{\mathrm{traj}} \ge 0$*, and hence*

$$\mathrm{EDL}^{\mathrm{warm}} \le \mathrm{EDL}^{\mathrm{cold}} + \Delta_{\mathrm{end}}. \tag{15}$$

If additionally $\Delta_{\mathrm{end}} \le 0$ *(warm start converges at least as well), then* $\mathrm{EDL}^{\mathrm{warm}} \le \mathrm{EDL}^{\mathrm{cold}}$.

(iii) ***Bounds on trajectory advantage.*** *The trajectory advantage satisfies*

$$n_{\mathrm{ft}} \cdot \min_{1 \le i \le n_{\mathrm{ft}}} \mathbb{E}[\Delta L_i] \le \Delta_{\mathrm{traj}} \le n_{\mathrm{ft}} \cdot \max_{1 \le i \le n_{\mathrm{ft}}} \mathbb{E}[\Delta L_i], \tag{16}$$

where $\Delta L_i \triangleq L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{cold}}) - L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{warm}})$.

Proof. By the expected area form (Theorem 1):

$$\mathrm{EDL}^{\mathrm{cold}} = \sum_{i=1}^{n_{\mathrm{ft}}} \mathbb{E}[L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{cold}})] - n_{\mathrm{ft}} \cdot \mathbb{E}[L_{\mathrm{ft}}(\theta_T^{\mathrm{cold}})], \tag{17}$$

$$\mathrm{EDL}^{\mathrm{warm}} = \sum_{i=1}^{n_{\mathrm{ft}}} \mathbb{E}[L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{warm}})] - n_{\mathrm{ft}} \cdot \mathbb{E}[L_{\mathrm{ft}}(\theta_T^{\mathrm{warm}})]. \tag{18}$$

Subtracting:

$$\begin{aligned} \mathrm{EDL}^{\mathrm{cold}} - \mathrm{EDL}^{\mathrm{warm}} &= \sum_{i=1}^{n_{\mathrm{ft}}} \mathbb{E}[L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{cold}}) - L_{\mathrm{ft}}(\theta_{i-1}^{\mathrm{warm}})] && (19)\\ &\quad - n_{\mathrm{ft}} \cdot \mathbb{E}[L_{\mathrm{ft}}(\theta_T^{\mathrm{cold}}) - L_{\mathrm{ft}}(\theta_T^{\mathrm{warm}})] && (20)\\ &= \Delta_{\mathrm{traj}} + n_{\mathrm{ft}} \cdot \mathbb{E}[L_{\mathrm{ft}}(\theta_T^{\mathrm{warm}}) - L_{\mathrm{ft}}(\theta_T^{\mathrm{cold}})], && (21) \end{aligned}$$

which establishes (13).

Part (i) follows immediately by setting $\Delta_{\mathrm{end}} = 0$.

For part (ii), trajectory dominance implies each summand in Δ_{traj} is non-negative, hence $\Delta_{\mathrm{traj}} \ge 0$. The stated inequality follows from rearranging (13).

For part (iii), note that $\Delta_{\mathrm{traj}} = \sum_{i=1}^{n_{\mathrm{ft}}} \mathbb{E}[\Delta L_i]$. Since the sum of n_{ft} terms is bounded below by n_{ft} times the minimum term and above by n_{ft} times the maximum term, the bounds follow. □

Remark 10 (Interpretation for transfer learning)**.** *Corollary 15 formalizes the intuition that pretraining amortizes learning: structure transferred from $\mathcal{D}_{\mathrm{pre}}$ reduces the information that must subsequently be extracted from S_{ft}. The fine-tuning EDL measures only the downstream* task-specific *information in S_{ft} not already encoded as inductive bias during pretraining. This*

yields a principled measure of transfer: the reduction in EDL attributable to pretraining measures how much relevant structure was transferred through the pretrained initialization.

In settings where $\mathcal{D}_{\mathrm{pre}}$ and $\mathcal{D}_{\mathrm{ft}}$ exhibit substantial shared structure, pretraining yields a correspondingly large reduction in fine-tuning EDL; conversely, when the two distributions share little structure, the reduction is negligible.

Remark 11 (When the simple approximation fails)**.** *A natural approximation suggested by Corollary 15 is*

$$\mathrm{EDL}^{\mathrm{cold}} - \mathrm{EDL}^{\mathrm{warm}} \approx n_{\mathrm{ft}} \cdot [L_{\mathrm{ft}}(\theta_0) - L_{\mathrm{ft}}(\theta_{T_{\mathrm{pre}}})], \tag{22}$$

corresponding to the upper bound in part (iii) evaluated at $i = 1$. This approximation is accurate when the gap ΔL_i between cold-start and warm-start loss remains approximately constant throughout the first epoch—i.e., when the training curves are parallel.

In practice, cold-start training often catches up to warm-start training as optimization progresses, causing $\Delta L_i \to 0$ as $i \to n_{\mathrm{ft}}$. In such cases, the trajectory advantage Δ_{traj} is substantially smaller than the bound (22) *suggests. The exact decomposition* (13) *captures this effect by integrating the instantaneous gap along the trajectory rather than extrapolating from initial conditions.*

Conversely, when the pretrained model lies in a qualitatively different loss basin that maintains its advantage throughout training, the approximation (22) *is tight.*

Remark 12 (Negative transfer)**.** *The decomposition also characterizes* negative transfer*: when pretraining on $\mathcal{D}_{\mathrm{pre}}$ harms performance on $\mathcal{D}_{\mathrm{ft}}$. If $L_{\mathrm{ft}}(\theta_{T_{\mathrm{pre}}}) > L_{\mathrm{ft}}(\theta_0)$—i.e., the pretrained initialization is* worse *than random for the downstream task—then $\Delta L_1 < 0$. Unless cold-start training degrades rapidly (unusual for well-designed optimizers), here we expect $\Delta_{\mathrm{traj}} < 0$ (negative transfer), indicating that pretraining* increased *the information that must be extracted from S_{ft}.*

This provides an information-theoretic diagnostic for negative transfer: compute the fine-tuning EDL from both initializations and compare.