Work in Progress: Information-theoretic characterization of the subregular hierarchy

Anonymous submission

Our goal is to link two different formal notions of complexity: the complexity classes defined by Formal Language Theory (FLT)—in particular, the Subregular Hierarchy (Heinz, 2018)—and Statistical Complexity Theory (Feldman and Crutchfield, 1998). The motivation for exploring this connection is that statistical complexity theory gives a highly general information-theoretic characterization of the use of memory resources during the generation and prediction of sequences, and factors involving memory resources have been hypothesized to explain why phonological processes seem to inhabit the Subregular Hierarchy (Lai, 2015). It is currently not known whether statistical complexity and FLT define equivalent complexity classes, or whether statistical complexity cross-cuts the usual FLT hierarchies. Our work begins to bridge the gap between FLT and information theory by presenting characterizations of certain subregular languages in terms of statistical complexity.

Statistical complexity theory. Statistical complexity theory deals with stochastic processes. For a process $X$ generating sequences of symbols indexed as $X_{t-2}, X_{t-1}, X_t, X_{t+1}, \ldots$, we define the notation $\overline{X}$ (“the past”) to mean $X_{t-2}, X_{t-1}$, and $\overline{X}$ (“the future”) to mean $X_t, X_{t+1}, \ldots$.

The statistical complexity of a stochastic process is the minimal amount of information about the past required to faithfully reproduce the future. If each symbol in the process is generated based on some memory representation $M$ of the past, we want to find the minimal information content of $M$ such that the generated symbols match the true process as closely as possible. Formally, the statistical complexity $S$ of a process $X$ is the entropy of $M$:

$$S \equiv \min_{M: D_{KL}[\overline{X}|\overline{X}|M]=0} H[M],$$

where $H[M]$ is the entropy of the discrete random variable $M$:

$$H[M] \equiv -\sum_x p_M(x) \log p_M(x),$$

and $D_{KL}[:,:,] \equiv$ is conditional Kullback-Leibler divergence (see Cover and Thomas, 2006).

Further insight comes from considering the different factors that contribute to statistical complexity. Using information-theoretic identities, we break the statistical complexity into two terms:

$$S = H[M] = I[M : \overline{X}] + H[M|\overline{X}] = \frac{I[X : \overline{X}]}{E} + \frac{H[M|\overline{X}]}{C},$$

where $I[:: :]$ is mutual information, the amount of information in one random variable about another. The term $E$ is called excess entropy and quantifies the amount of information in the past which is useful for predicting the future. The term $C$ is called crypticity and quantifies the amount of information stored in $M$ which does not end up being useful for predicting the future.

These quantities are easily understood in terms of memory resources used for incremental language production and comprehension. Excess entropy measures integration cost: it says how many bits of information from the past are used when processing the future. Statistical complexity measures memory load or storage cost; it can be finite even for non-finite-state processes, as long as the sum in Eq. 1 converges.

Preliminary results. We study subregular languages defined using Probabilistic Deterministic Finite-State Automata (PDFA). A PDFA is characterized by a set of internal states $Q$, an alphabet $\Sigma$, a probability distribution of symbols $\Sigma \rightarrow Q$ conditional on a state $\in Q$, a function $Q \times \Sigma \rightarrow Q$ defining which state the machine transitions into after emitting a symbol, and distinguished initial
and final states. Our indexing convention is: at time $t$, the PDFA is in state $q_t$; it generates symbol $x_t$ before transitioning into the next state $q_{t+1}$. We use the following construction to generate a stationary ergodic stochastic process from a PDFA: upon entering the final state, the PDFA emits an end-of-word symbol # and transitions back into the initial state. The resulting infinite stream of symbols is amenable to analysis using statistical complexity theory.

Below, we describe how to calculate $S$, $E$, and $C$ from the minimal PDFA for Strictly $k$-Local (SL$_k$) languages. Table 1 shows calculated statistical complexity, excess entropy, and crypticity for the PDFAs of SL$_k$. Locally $k$-Testable (LT$_k$), and Locally $k$-Threshold Testable (LTT$_k$) classes when $k = 2$.

![Figure 1: SL$_2$ PDFA of $*ab$, $\Sigma = \{a, b\}$](image1)

In the case of SL$_2$ languages, we compute $E$ by constructing a symbol transition matrix, a stochastic matrix whose entries represent the probability of going into state $q_{t+1}$ after being in state $q_t$: $p(q_{t+1}|q_t) = \sum_{x_t \in \Sigma} p(x_t|q_t)p(q_{t+1}|x_t,q_t)$. Then the distribution $Q$ is given by the left eigenvector of the state transition matrix associated with eigenvalue 1.

For an SL$_k$ language, statistical complexity is upper bounded as $S \leq (k-1) \log |\Sigma|$.

**Excess entropy.** For SL$_k$ languages,

$E = I[X_{t-k+2}, \ldots, X_{t-1} : X_t, X_{t+k-2}]$.

In the case of SL$_2$ languages, we compute $E$ by constructing a symbol transition matrix, a stochastic matrix whose entries represent $p(x_{t+1}|x_t)$, marginalizing over $q_t$ and $q_{t+1}$. We also need the stationary distribution over symbols, derived from the symbol transition matrix by the same procedure as above.

**Crypticity.** Crypticity $C = S - E$. In general, crypticity is bounded above by the uncertainty about the emitting state given a symbol:

$C \leq H[Q_t|X_t]$,

with equality iff $X$ is an SL$_2$ language.

### Table 1: Information quantities for PDFAs shown in figures. SL$_3$ = Figure 1; LT$_2$ = Figure 2; LTT$_2$ = Figure 3. Quantities marked with $\leq$ or $\geq$ are bounds based on Markov approximations.

|          | SL$_2$ | LT$_2$ | LTT$_2$ |
|----------|--------|--------|---------|
| Statistical complexity | 0.97   | 1.58   | 2.08    |
| Excess entropy    | $\geq 0.78$ | $\geq 1.13$ |          |
| Crypticity        | $\leq 0.80$ | $\leq 0.96$ |          |
| $E/S$ ratio       | 0.25   | $\geq 0.49$ | $\geq 0.54$ |

**Statistical complexity.** For a PDFA which has the minimal number of states, the minimal-entropy memory representation $M$ is simply the internal state itself (Travers and Crutchfield, 2011). Therefore the statistical complexity $S = H[Q]$ where $Q$ is the stationary distribution over states. To get the distribution $Q$, we first construct a state transition matrix: a stochastic matrix whose

**References**

- Thomas M. Cover and J.A. Thomas. 2006. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.
- David P. Feldman and James P. Crutchfield. 1998. Measures of statistical complexity: Why? *Physica Letters A*, 238(4-5):244–252.
- Jeffrey Heinz. 2018. The computational nature of phonological generalizations. *Phonology, Phonetics and Phonology*, pages 126–195.
- Regine Lai. 2015. Learnable vs. unlearnable harmony patterns. *Linguistic Inquiry*, 46(3):425–451.
- Nicholas F. Travers and James P. Crutchfield. 2011. Asymptotic synchronization for finite-state sources. *Journal of Statistical Physics*, 145(5):1202–1223.