Many large datasets in research, archives, ML training, etc.

Consider a sequence \((X_1, X_2, \ldots, X_n)\) with \(X_1, X_2, \ldots, X_n \sim P_X\).

Given a random variable over an alphabet \(\mathcal{X}\), a
lossless source code (l.s.c.) consists of functions
\(f : \mathcal{X} \rightarrow \{0, 1\}^*\) and \(g : \{0, 1\}^* \rightarrow \mathcal{X}\)
such that \(g(f(x)) = x\).

To encode symbols with few bits, we usually minimize the average code word length:
\[M^*(P_X) \triangleq \min \{E[l(f(X))] \mid (f, g) \text{ is a prefix-free l.s.c.}\}\]

Traditional compression algorithms operate on stream data.
If we don’t need to preserve order on elements, we can save space and bandwidth with a “dataset” compression algorithm.

Consider a “dataset”, i.e. a set of samples where order doesn’t matter: \(\{X_1, X_2, \ldots, X_n\}\).

Define a lossless dataset source code (l.d.s.c.) \(f : \mathcal{X} \rightarrow \{0, 1\}^*\) and \(g : \{0, 1\}^* \rightarrow \mathcal{X}\) such that
\[g(f(x^n)) = \pi \circ x^n \text{ for all } x^n \in \mathcal{X}^n \text{ where } \pi \text{ is a permutation.}\]

For a l.d.s.c. we minimize
\[M^*_d(P_X) \triangleq \min \{E[l(f(X^n))] \mid (f, g) \text{ is a prefix-free l.d.s.c.}\}\]

Note: dataset code length \(\leq\) sequence code length

Motivation

Traditional compression algorithms operate on stream data. If we don’t need to preserve order on elements, we can save space and bandwidth with a “dataset” compression algorithm.

Idea: in many data, the “features” are not truly independent from each other. For example, pixels in an image.

Predictive coding: instead of encoding every value, can encode some values, a model to obtain adjacent ones, and the “error” between predictions and true value.

Combining ideas from Theorem 1 and JPEG-LS:

1. Build nearest neighbors graph of dataset
2. Obtain reordering by traversing a MST
3. Train predictor on context from within images and adjacent images
4. Entropy coding

Theoretical Results

Theorem 1 (l.d.s.c. via data structures):
Let \(\hat{X}^n = \pi \circ X^n\) where \(\pi\) is a permutation drawn uniformly at random. Moreover, let \(S\) be such that
\[1) \ S \rightarrow X^n \rightarrow \hat{X}^n \text{ and } X^n \rightarrow S \rightarrow \hat{X}^n\]
\[2) \ H(S|X^n) = H(S|\hat{X}^n) = 0\]

Then \(M^*_d(P_S) = M^*_n(P_X)\),
\[H(S) \leq M^*_n(P_X) < H(S) + 1\]
and \(H(S) = I(X^n; \hat{X}^n) \leq \min\{|\mathcal{X}| \log_2(n+1), n \log_2 |\mathcal{X}|\}\)

Experiments

Idea: in many data, the “features” are not truly independent from each other. For example, pixels in an image.

Predictive coding: instead of encoding every value, can encode some values, a model to obtain adjacent ones, and the “error” between predictions and true value.

Combining ideas from Theorem 1 and JPEG-LS:

1. Build nearest neighbors graph of dataset
2. Obtain reordering by traversing a MST
3. Train predictor on context from within images and adjacent images
4. Entropy coding

Benchmarks

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Ordering</th>
<th>Context strategy</th>
<th>Final size (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>Logreg</td>
<td>Random</td>
<td>112.41</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>Logreg</td>
<td>Random</td>
<td>112.41</td>
</tr>
</tbody>
</table>

Madeleine Barowsky, Alexander Mariona, and Flavio du Pin Calmon
Harvard University