Large Alphabet Problem

- **Universal Compression**: In most modern universal lossless compression algorithms, the statistics of data is captured by using some forms of context memory to learn the statistics of the past data and then to use this information to compress new data.
- **Large Alphabet Problem**: Statistical learning for the case of large alphabet data is challenging, mainly due to huge amount of model parameters and also the need for massive number of training samples. The alphabet size in modern seismic data is $2^{12}$ which makes the traditional approaches impractical specially for smart sensor applications.
- We propose two approaches to tackle the large alphabet problem:

1. **Lossless compression**: using a modified version of CTW algorithm combined with memory assisted compression. To cope with non-stationarity of the seismic data we also incorporate MDL clustering into our compression framework.
2. **Near lossless compression**: We use the idea behind oversampled one-bit A/D conversion as a way to translate the resolution in the amplitude to the resolution in time by creating a sequence of non-uniform binary samples. This method would significantly reduce the dynamic range at the cost of increasing the number of samples.

CTWL + Memory Assisted Compression

- As a modification to the well known CTW algorithm, we use a reduced context to capture the statistics of past data to make the context memory learning possible for seismic data.

To cope with the non-stationary behavior of seismic traces, we employ Minimum Description Length (MDL) principle to partition the memory to several clusters. In the training phase, we classify short sequences of trace data in the memory using an MDL based classifier and in the online compression we simply use the pre-trained clustered memory to compress the new data while the model at both encoder and decoder will be updated by each new data transmitted.

Regime Change: Time Resolution Refinement vs Amplitude Quantization Refinement

Using an optimally designed non-linear reconstruction, the rate distortion characteristics attained by quantization refinement is almost achievable by using a 1-bit quantized sequence of non-uniform samplings.

- **Quantization R/D characteristic**:
  $$E((f(t) - f_q(t))^2) = O(2^{-2B})$$
- **1-bit oversampled R/D characteristics**:
  $$E((f(t) - f_q(t))^2) = a_1 O(2^{-aB})$$

Generating 1 sequence of 1-bit non-uniform using a sawtooth dither function

Experimental Results

This figure shows the rate/distortion characteristics for the proposed method using a saw tooth dither with two different dither frequency ratios compared to the results achieved by the traditional sinusoid dither as well as the rate/distortion for DCT.

Publicly available USGS seismic data was used in this experiment

Conclusion

- We investigated two different approaches to compress the large alphabet seismic data for two cases of lossless and near lossless compression.
- In the first method, we develop a modified CTW algorithm suitable for lossless compression of large alphabet.
- In the second method, we develop a near lossless scheme by changing the regime and using zero-crossings of the oversampled signal.
- We also introduce an MDL based classifier to increase the performance of the system.