

Relation between Fractal Dimension and Performance of Vector Quantization

Krishna Kumaraswamy¹, Christos Faloutsos¹,
Guoqiang Shan² and Vasileios Megalooikonomou²

¹Center For Automated Learning and Discovery, CMU

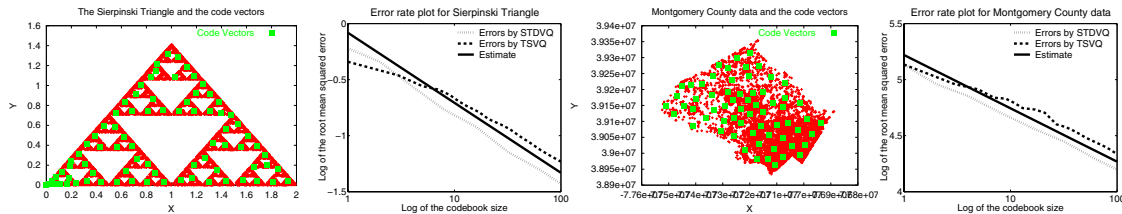
²Department of Computer and Information Sciences, Temple University

We show that the performance of a vector quantizer is related to the intrinsic (“fractal”) dimension of the data set. For a perfectly self-similar object with r similar pieces, each scaled down by a factor s the fractal dimension \mathcal{D} is defined as $\mathcal{D} = \frac{\log r}{\log s}$. For a given data set, we can measure the intrinsic (fractal) dimension as the slope of the *correlation integral* defined as: $C(r) = \#(\text{pairs within distance } r \text{ or less})$.

For a perfectly self-similar data set of size N we show that the error $E(k)$ is related to the fractal dimension \mathcal{D} as follows:

$$\log E(k) = c - \frac{1}{\mathcal{D}} \cdot \log k$$

where k is the number of codewords (representative elements) for Vector Quantization, $E(k)$ the root mean squared error (RMSE) and c is a constant given by $c = \log \sum_{i=1}^N (x_i - \bar{x})^2$ which is the log error when the rate is zero (i.e., we do not transmit anything) and we use the mean value as the best representative of all points. The proof is omitted due to space limitations. Our conjecture is that for a statistically self-similar data set, the same relation holds where the fractal dimension has the same definition constrained to a range of scales (r_{min}, r_{max}). We performed experiments to confirm our result on synthetic and real data sets. The fitted line for the error-rate plot is very close to the line estimated using our result (see figure). To further verify our result, we computed the slope and compare it to the estimate of the fractal dimension obtained using the correlation integral.



From a practical point of view, our results help us estimate the optimal performance of any Vector Quantizer. Moreover, the computation of the correlation fractal dimension is linear on the number of data points and significantly faster (about 10x) than vector quantization itself. Previous related work includes that of Zador [1] who does not prove a general result for fractals but he does show that the usual high rate formula for absolutely continuous densities applies to a case with no absolutely continuous components with the usual Euclidean dimension replaced by the fractal dimension.

Acknowledgment: This work was supported in part by NSF (IIS-0237921).

References: [1] P.L. Zador, “Asymptotic quantization error of continuous signals and the quantization dimension,” IEEE Trans. Inform. Theory, vol. 28, pp. 139-148, March 1982.