Rates of Convergence in the Source Coding Theorem, in Empirical Quantizer Design, and in Universal Lossy Source Coding

Tamás Linder[1] Gábor Lugosi[2] Kenneth Zeger[3]

[1]Dept. of Telecommunications, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary.
[2]Dept. of Mathematics, Faculty of Elect. Eng., Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary.
[3]Coordinated Science Laboratory, Dept. of Elect. and Comp. Engineering, University of Illinois, Urbana, IL 61801

Abstract — **Rates of convergence results are established for vector quantization. Convergence rates are given for an increasing vector dimension and/or an increasing training set size. In particular, the following results are shown for memoryless real valued sources with bounded support at transmission rate R: (1) If a vector quantizer with fixed dimension k is designed to minimize the empirical MSE with respect to m training vectors, then its MSE for the true source converges almost surely to the minimum possible MSE as $O(\sqrt{\log m/m})$; (2) The MSE of an optimal k-dimensional vector quantizer for the true source converges, as the dimension grows, to the distortion-rate function $D(R)$ as $O(\sqrt{\log k/k})$; (3) There exists a fixed rate universal lossy source coding scheme whose per letter MSE on n real valued source samples converges almost surely to the distortion-rate function $D(R)$ as $O(\sqrt{\log\log n/\log n})$; and (4) Consider a training set of n real valued source samples blocked into vectors of dimension k, and a k-dimensional vector quantizer designed to minimize the empirical MSE with respect to the $m = \lfloor n/k \rfloor$ training vectors. Then the MSE of this quantizer for the true source converges almost surely to the distortion-rate function $D(R)$ as $O(\sqrt{\log\log n/\log n})$, if one chooses $k = \lfloor \frac{1}{R}(1-\epsilon)(\log n)\rfloor$ $\forall\epsilon \in (0,1)$.**

Let $Q_{N,k}$ denote a k dimensional, N level nearest neighbor vector quantizer. Let $Z, Z_1, \ldots, Z_m \in \mathcal{R}^k$ be independent identically distributed random vectors (training data) and define the *average distortion* (mean square) as $\Delta(Q_{N,k}) = E\|Z - Q_{N,k}(Z)\|^2$ and its *empirical distortion* as $\Delta_m(Q_{N,k}) = \frac{1}{m}\sum_{i=1}^m \|Z_i - Q_{N,k}(Z_i)\|^2$.

Theorem 1 *Let $Z_1, Z_2, \ldots \in \mathcal{R}^k$ be i.i.d. random vectors such that $\Pr\{\|Z_1\|^2 \le B\} = 1$ and $m(t/8B)^2 \ge 2$. Suppose an N-level, k-dimensional quantizer, $Q^*_{m,N,k}$ is designed to minimize the empirical MSE over a training set of m vectors $Z_1, \ldots, Z_m$. Then the difference between the MSE of this quantizer for the true source and that of the best quantizer, for the true source, satisfies*

$$\Pr\{\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) > t\} \le 4(2m)^{N(k+1)} e^{-mt^2/(512B^2)}. \quad (1)$$

Corollary 1 *Let $Z_1, Z_2, \ldots \in \mathcal{R}^k$ be an i.i.d. source that is bounded with probability one and suppose an N-level, k-dimensional quantizer, $Q^*_{m,N,k}$, is designed to minimize the empirical MSE over a training set of m vectors $Z_1, \ldots, Z_m$. Then its MSE for the true source converges almost surely as $m \to \infty$ to the minimum MSE of the best quantizer, $Q^*_{N,k}$, for the true source, at a rate*

$$\Delta(Q^*_{m,N,k}) - \Delta(Q^*_{N,k}) = O\left(\sqrt{\frac{\log m}{m}}\right) \quad a.s.$$

Theorem 2 *Let $X_1, X_2, \ldots$ be a real valued i.i.d. source that is bounded with probability one and has distortion-rate function $D(R)$. Then for every $R > 0$ with $D(R) > 0$ there is a constant c such that for every k the difference between the per letter MSE of the best k-dimensional quantizer of rate R and $D(R)$ satisfies*

$$D_k(R) - D(R) \le c\sqrt{\frac{\log k}{k}}.$$

Definition 1 *For $R > 0$ a sequence of pairs of functions (f_n, ϕ_n) of the form*

$$f_n : \mathcal{R}^n \to \{0,1\}^{\lfloor nR \rfloor} \text{ and } \phi_n : \{0,1\}^{\lfloor nR \rfloor} \to \mathcal{R}^n$$

is called an **almost sure universal source coding scheme** *of rate R with respect to a family of real sources, if for each source $X_1, X_2, \ldots$ in the family, the n-blocks $(Y_{1,n}, \ldots, Y_{n,n}) = \phi_n(f_n(X_1, \ldots, X_n))$ satisify ($D(R)$ is the distortion-rate function of the source) $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n (X_i - Y_{i,n})^2 = D(R)$ a.s.*

Theorem 3 *For every rate $R > 0$ there exists an almost sure universal source coding scheme for the family of stationary and ergodic real sources with finite second moment. Moreover, if $D(R)$ is the distortion-rate function of any source in the subfamily of i.i.d real-valued sources with bounded support, then for every $R > 0$ such that $D(R) > 0$, there is a constant $c > 0$ such that the difference between the per letter sample MSE and $D(R)$ decays $\forall\epsilon \in (0, 1/2)$ as*

$$\frac{1}{n}\sum_{i=1}^n (X_i - Y_{i,n})^2 - D(R) \le c\sqrt{\frac{\log\log n}{\log n}} + o\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2}-\epsilon}\right) \quad a.s.$$

Theorem 4 *Let $X_1, \ldots, X_n$ be n samples from a real valued i.i.d. source that are bounded with probability one, and suppose these samples are blocked into k-dimensional "training" vectors $Z_1, \ldots, Z_m$, where $Z_i = (X_{(i-1)k+1}, \ldots, X_{ki})$ and $m = \lfloor \frac{n}{k} \rfloor$. Let $Q^*_{m,N,k}$ be a k-dimensional vector quantizer designed to minimize the empirical MSE for the m training vectors. Then by choosing $k = \lfloor \frac{1}{R}(1-\epsilon)\log n \rfloor$, for any $\epsilon \in (0,1)$, the per letter MSE of $Q^*_{m,N,k}$, for the true source, converges to the distortion-rate function at the rate*

$$\frac{1}{k}\Delta(Q^*_{m,N,k}) - D(R) = O\left(\sqrt{\frac{\log\log n}{\log n}}\right) \quad a.s. \quad (2)$$

[1]Supported in part by NSF Grant No. NCR-92-96231.