A Type Generator Model for Adaptive Lossy Compression

Ram Zamir [1] and Kenneth Rose [2]

[1] Dept. of EE - Systems, Tel Aviv University, Tel Aviv 69978, ISRAEL *zamir@eng.tau.ac.il*
[2] Dept. of ECE, University of California, Santa Barbara, CA 93106 *rose@ece.ucsb.edu*

Abstract — **We propose a random coding model which gives insight into the underlying mechanism of lossy string-matching algorithms. This idealized model is based on manipulating types rather than codewords. We model the adaptive codebook as a mixed type random code, whose type distribution evolves with time as its dimension increases. The model behavior exhibits a mechanism of *natural type selection* which reinforces "good types."**

I. The lossy string-matching exponent

Consider matching an individual sourceword $\underline{x} \in \mathcal{X}^n$ with a random vector $\underline{Y} \in \mathcal{Y}^n$ uniformly distributed over the type class $T(q)$. Here q is an arbitrary type in $\mathcal{Q}_n$ = the set of possible types of vectors in $\mathcal{Y}^n$, and a "match" is the event $\rho(\underline{x}, \underline{Y}) \leq D$. We obtain the asymptotic result:

$$-\frac{1}{n}\log \Pr[\rho(\underline{x}, \underline{Y}) \leq d] = I_m(p\|q,d) + o(1)\ , \qquad (1)$$

where $o(1)$ goes to zero as $n \to \infty$. The quantity $I_m(p\|q,d)$ in (1) is defined as

$$I_m(p\|q,d) = \min_{W \in \mathcal{W}_{p,q,d}} I_W(X;Y)\ , \qquad (2)$$

where $I_W(X;Y)$ is the mutual information induced by the joint distribution W, and $\mathcal{W}_{p,q,d} = \{W : \sum_{y\in\mathcal{Y}} W(x,y) = p(x), \sum_{x\in\mathcal{X}} W(x,y) = q(y), \sum_{(x,y)\in(\mathcal{X},\mathcal{Y})} W(x,y)\rho(x,y) \leq d\}$. If $\mathcal{W}_{p,q,d}$ is empty, we define $I_m = \infty$. (See also [4, sec.III-C].)

II. Matching with a Mixed-Type Code

Consider next a mixed type random code having *type-spectrum* $w(q)$, i.e., a random code of length n, whose codwords are generated according to the distribution $w(q)$ over the types in $\mathcal{Q}_n$. The minimum possible rate required by this code to ensure distortion d for a stationary source with marginal distribution p, is given for large n by

$$R_n(w;p,d) = \min_{q'\in\mathcal{Q}_n}\left\{I_m(p\|q',d) + \frac{1}{n}\log(1/w(q'))\right\}\ . \qquad (3)$$

Notable special cases: *i)* the "pure-q-type" random code, $w(q') = \delta(q-q')$ where $R_n(w;p,d) = I_m(p\|q,d)$; *ii)* the "white spectrum" random code, $w(q') = 1/|\mathcal{Q}_n|$ where Since $|\mathcal{Q}_n| = O(n^{|\mathcal{Y}|})$ we get $R_n(w;p,d) = \min_{q'\in\mathcal{Q}_n} I_m(p\|q',d) \to R(p,d)$; *iii)* a code generated *i.i.d.* $\sim q$, where

$$R_n(w;p,d) \to \min_{q'}\left\{I_m(p\|q',d) + \mathcal{D}(q'\|q)\right\} \qquad (4)$$

and $\mathcal{D}(\cdot\|\cdot)$ is the divergence function (This is closely related to the independent result in [2]); *iv)* the special case of (iii) where $p = q$ corresponds to the algorithm of [1], and the rate is $R = \min_{q'}\{I_m(p\|q',d) + \mathcal{D}(q'\|p)\}$.

Suppose that for each source vector $\underline{x}$, the code generates random codwords according to the type spectrum $w(q)$ until it finds a codword that d-matches $\underline{x}$. The *a posteriori type distribution* of a mixed type random code is defined as the conditional probability that the d-matching codword has type q, given that there was a match.

III. Natural Selection of Types

In the model we propose, the mixed type random code above plays the role of the temporal codebook in an adaptive compression algorithm. Following the tree structure of Lempel Ziv's codebook, the adaptive compression model increments the code dimension at each time step by growing the types in a treelike fashion. The evolution of the type spectrum is given recursively by

$$w_{n+1}(q) = \frac{1}{|\mathcal{Y}|}\sum_{q'\in\mathcal{A}(q)} P_n^{ap}(q')\ ,\ \ q \in \mathcal{Q}_{n+1}, \qquad (5)$$

where $\mathcal{A}(q)$ denotes the set of n-types from which the $n+1$-type q descends by adding one symbol, and $P_n^{ap}(\cdot)$ denotes the a posteriori type distribution of the code at time step n.

Let $q^* = \arg\min_q I_m(p\|q,d)$ denote the optimal reconstruction distribution for a source $\sim p$ at distortion d. Our main result states that the type spectrum in the adaptive compression model above collapses asymptotically on q^*, i.e., $w_n(q) \Rightarrow \delta(q-q^*)$ as $n \to \infty$ where $\Rightarrow$ denotes weak convergence. We thus discern a "Darwinian mechanism" of natural selection whereby, as time passes, only the "fittest" types survive. Consequentially, we have by (3) that the minimum rate of the code goes to the rate distortion function of the source $R(p,d)$ as n goes to infinity.

Beyond the theoretical interest, this model guides the development of a family of practical adaptive compression algorithms [3] which are based on "tree-code matching" and naturally extend the Lempel-Ziv algorithm to lossy coding.

References

[1] Y. Steinberg and M. Gutman, "An algorithm for source coding subject to a fidelity criterion based on string matching," *IEEE Trans. Inform. Theory*, IT-39, pp. 877–886, 1993.

[2] E.-H Yang and J. Kieffer, "On the performance of data compression algorithms based upon string matching," *IEEE Trans. Information Theory*, preprint.

[3] R. Zamir and K. Rose, "Towards lossy Lempel-Ziv: natural type selection," *Proc. of the Inform. Theory Workshop, Haifa, Israel*, p. 58, June 1996.

[4] Z. Zhang and V. Wei, "An on-line universal lossy data compression algorithm via continuous codebook refinement-part I: Basic results," *IEEE Trans. Inform. Theory*, IT-42, pp. 803–821, 1996.

[2]This work was supported in part by the NSF under grant no. NCR-9314335, the University of California MICRO program, ACT Networks, Advanced Computer Communications, Cisco Systems, DSP Group, DSP Software Engineering, Fujitsu, General Electric Companuy, Hughes Electronics, Intel, Nokia Mobile Phones, Qualcomm, Rockwell International, and Texas Instruments.