Alternate Learning and Compression approaching $R(D)$

Ram Zamir (TAU) and Kenneth Rose (UCSB)

May 2024

Extended Abstract

The inherent trade-off in on-line learning is between exploration and exploitation. A good balance between these two (conflicting) goals can achieve a better long-term performance. Can we define an optimal balance? We propose to study this question through a backward-adaptive lossy compression system, which exhibits a "natural" trade-off between exploration and exploitation.

From a rate-distortion perspective, a universal lossy compression system needs to learn the rate-distortion achieving distribution, Q^*, in order to construct an optimal codebook. This is unlike the lossless case, where strings of the source itself have the "right" distribution and can serve as a basis for the code dictionary (as done, e.g., in Tunstall or Lempel Ziv coding [23]). When the source distribution is known, the system designer can use the well known Arimoto-Blahut algorithm to compute Q^* [1, 4]. Less known is the fact that even if the source distribution is unknown, a sequential scheme can learn Q^* on the fly, while compressing consecutive source vectors.

Such an asymptotically optimal sequential coding scheme, called Natural Type Selection (NTS), was described in [22]. It is based on the observation that the empirical distribution of the d-matching codeword is atypical with respect to, and better than, the codebook generating distribution. Hence, this empirical distribution provides information on how to adjust the codebook generating distribution in the direction of Q^*.

Specifically, NTS-based encoding alternates between two phases:

1. *Compression phase* - where the encoder finds the index of the first d-matching codeword and transmits it to the decoder; and

2. *Learning phase* - where the encoder and decoder estimate the type (or some other representative parameter) of the d-matching codeword.

After these two phases, the codebook is updated (identically) by the encoder and the decoder. As it turns out, [22], in the limit of a large word length L, the compression and learning phases above stochastically simulate an iteration

of the Blahut algorithm for RDF computation, and hence converge to the RDF; see Arimoto [1] and Blahut [4].[1]

The NTS mechanism was extended and investigated from various information theoretic perspectives and settings, [13, 8, 9, 17, 16, 14, 18]. In this extended abstract we propose to consider its implications within the context of on-line learning and reinforcement learning, e.g. [2]. This is a preliminary study, and we do not formally establish new results. Rather, we propose a fresh look that we believe may be of some interest to researchers in the intersection of compression and learning.

Why we need to explore?

Exploration is *not* intrinsically inherent to universal compression, but rather a consequence of a *backward*-adaptive model of such system. Let us clarify the difference between the two modes of adaptation: forward and backward. In forward ("batch" / "two-part") adaptation, e.g., dynamic Huffman or CELP speech coding, the encoder learns the source statistics, computes the optimal encoding parameters, and sends them to the decoder as a header ("side information"), before it begins to encode the source data. In backward ("sequential") adaptation, e.g., Lempel-Ziv or ADPCM, both the encoder and decoder learn the parameters from past reconstructed samples, so there is no explicit transmission of side information. See [12, 10, 11]. While in lossless compression the two modes of operation are essentially equivalent (e.g. [20]), in the lossy case they are fundamentally different: forward adaptation learns from the *clean* version of the source, while backward adaptation learns from the *noisy* (quantized) version. Furthermore, the difference between the two grows in significance with increase in the prescribed distortion level. In information-theoretic terms, as D increases from 0 to $D_{\max}$, the reconstruction distribution Q^*, which achieves the rate-distortion bound, deviates from the source distribution P, progressively concentrates on a smaller subset of the reconstruction alphabet, and eventually at $D = D_{\max}$ collapses to a single probability mass point at the "centroid" letter.

We argue that for a memoryless source and a given (mismatched) reconstruction codebook, the type Q of the reconstruction sequence is a *sufficient statistic* for learning Q^* in a backward mode. Furthermore, at large distortion (D close to $D_{\max}$), Q carries almost no information about the source distribution P itself, and therefore Q^* cannot be computed directly from Q. Thus, a type's goodness (for compression) can only be established when a codeword of this type d-matches a source word. We thus conclude that in backward-adaptive lossy compression at high distortion, explicit exploration of types is necessary in order to find Q^*.

[1] This description amounts to a fixed-distortion variant of Blahut [6, 7]. Replacing d-match by a weighted distortion-code-length sum amounts to the usual "fixed-slope" version of the Blahut algorithm [4].

Rates of convergence

To study the efficiency of NTS, we first consider the speed of convergence of several related learning algorithms.

The convergence of the Blahut algorithm to the RDF is of the order of $O(1/N)$ after N iterations. This was shown in ([5, 7]) by writing the sum of the gaps of the intermediate rates $R(P, Q_N, D)$ from $R(P, D)$ as a telescopic sum that is bounded by a finite constant = the divergence between the optimum output distribution Q^* and the initial output distribution Q_0. Thus, the gaps must decrease at least as fast as $O(1/N)$. A similar decrease is believed to hold for iterative design of K-level quantizers, via the Lloyd-Max algorithm (alternations between centroids computation and thresholds computation). While faster-than-Blahut computation of $R(D)$ is possible at small distortions due to the known structure of the RDF (a Lagrangian solution that assumes that all the output alphabet letters have positive probability), at high distortion most letters are inactive, so the situation is similar to computation of the optmal 2-level quantizer.

As for the effect of the word length, universal compression schemes, lossy and lossless, are known to exhibit redundancy on the order of $O(\log(L)/L)$. In the lossy case, this is due to finite vector-quantization loss ("granular gain"), as well as to the cost of universality (statistical learning).

Exploration strategies

The two-phase compression-learning mechanism of NTS can be viewed as an exploitation step (compression), followed by an exploration step (learning). The amount of exploration is governed by the frequency of atypical codewords in the random codebook. In a codebook generated i.i.d, this frequency is known to decay exponentially with the divergence between the codeword type and the codebook generating distribution. This divergence also governs the steps of the corresponding Blahut algorithm, and dictates its rate of convergence.

Is this "natural" trade-off between exploration and exploitation optimal?

In fact, even if the source is memoryless, a codebook distribution that is not i.i.d. can emphasize rare types and accelerate the movement towards Q^*. A simple example is a uniform weighting of all type classes [6, 21]. More generally, to obtain a richer codebook distribution, one may use a mixture over a parametric family of distributions, as done in the universal lossless case, e.g., in Minimum Description Length [3] or Context-Tree Weighting [19, 20, 15].

This view suggests a tradeoff between "breadth and depth", i.e., wide exploration versus narrow exploration: a richer universal mixture implies higher probability of rare types and reduced probability of typical codewords. This may be disadvantageous as we get closer to Q^*, due to the inherent cost in coding rate (especially in the non-asymptotic regime where the word length L is moderate). Hence, there should be an optimal schedule for narrowing the richness of the universal mixture as the NTS gets closer to Q^*.

References

[1] S. Arimoto. An algorithm for calculating the capacity of an arbitrary discrete memoryless channel. *IEEE Transactions on Information Theory*, IT-18:14–20, 1972.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

[3] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

[4] R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, IT-18:460–473, 1972.

[5] P. Boukris. An upper bound on the speed of convergence of the blahut algorithm for computing rate-distortion functions (corresp.). *IEEE Transactions on Information Theory*, 19(5):708–709, 1973.

[6] T. M. Cover and J. A. Thomas. *Elements of Inf. Theory.* Wiley-Interscience, 2006.

[7] I. Csiszar and G. Tusnady. Information geometry and alternating minimization procedures. *Statist. Decision*, (1):205–237, 1984.

[8] A. Elshafiy, M. Namazi, and K. Rose. On effective stochastic mechanisms for on-the-fly codebook regeneration. In *IEEE International Symposium on Inf. Theory (ISIT)*, 2020.

[9] A. Elshafiy, M. Namazi, R. Zamir, and K. Rose. On-the-fly stochastic codebook re-generation for sources with memory. In *IEEE Information Theory Workshop (ITW)*, 2021.

[10] A. Gersho and R. M. Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 1992.

[11] J. D. Gibson. *Digital compression for multimedia: principles and standards.* Morgan Kaufmann, 1998.

[12] N. S. Jayant and P. Noll. *Digital coding of waveforms: principles and applications to speech and video*, volume 2. Prentice-Hall Englewood Cliffs, NJ, 1984.

[13] Y. Kochman and R. Zamir. Adaptive parametric vector quantization by natural type selection. In *Data Compression Conference (DCC)*, 2002.

[14] Y. Kochman and R. Zamir. Computation of the rate-distortion function relative to a parametric class of reproductions. In *PROCEEDINGS OF THE ANNUAL ALLERTON CONFERENCE ON COMMUNICATION CONTROL AND COMPUTING. Vol. 41. No. 1.*, 2003.

[15] N. Merhav and M. Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Transactions on Information Theory*, 41(3):714–722, 1995.

[16] D. Modha and N. Santhanam. Making the correct mistakes. In *Data Compression Conference (DCC'06)*, pages 302–311, 2006.

[17] H. Ratson and R. Zamir. Rate-distortion in non-convex families. In *2023 IEEE Information Theory Workshop (ITW)*, pages 87–91, 2023.

[18] S. Tridenski and R. Zamir. Channel input adaptation via natural type selection. *IEEE Transactions on Information Theory*, 66(4):2078–2090, 2020.

[19] F. Willems, Y. Shtarkov, and T. Tjalkens. The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995.

[20] F. Willems, Y. Shtarkov, and T. Tjalkens. Reflections on "the context tree weighting method: Basic properties". *Newsletter of the IEEE Information Theory Society*, 47(1), 1997.

[21] R. Zamir and K. Rose. A type generator model for adaptive lossy compression. In *Proceedings of IEEE International Symposium on Information Theory*, pages 186–, 1997.

[22] R. Zamir and K. Rose. Natural type selection in adaptive lossy compression. *IEEE Trans. on Inf. Theory*, 47:99–110, 2001.

[23] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. on Inf. Theory*, 23(3):337–343, 1977.