Locally Decodable Codes

Sergey Yekhanin

Microsoft Research Silicon Valley
yekhanin@microsoft.com

An r-query Locally Decodable Code (LDC) C encodes k-symbol messages $\mathbf{x}$ to n-symbol codewords $C(x)$. For every $i \in [k]$ the decoder has a family $\mathcal{D}_i$ of r-subsets of the universe $[n]$. These families are carefully designed to ensure that for all messages $\mathbf{x}$, for each $i \in [k]$ and $S \in \mathcal{D}_i$ the symbol $\mathbf{x}(i)$ can be recovered from accessing r coordinates of $C(\mathbf{x})$ that belong to S. Furthermore families $\mathcal{D}_i$ are required to satisfy the following two conditions:

- *Smoothness.* For each $i \in [k]$, the r-tuples in $\mathcal{D}_i$ cover the universe $[n]$ uniformly, i.e., each $j \in [n]$ belongs to the same number of sets in $\mathcal{D}_i$;
- *Error tolerance.* If the decoder for the i-th message symbol is invoked on a codeword $C(\mathbf{x})$ that has been adversarially corrupted in a small fraction of coordinates; then all but a small fraction of sets in $\mathcal{D}_i$ still yield the correct value of $\mathbf{x}(i)$.

The classical Hadamard code encoding k-bit messages to $n = 2^k$-bit codewords provides the simplest nontrivial example of locally decodable codes. Here $r = 2$. Every coordinate in the Hadamard code corresponds to one of 2^k subsets of $[k]$ and stores the XOR of the corresponding bits of the message $\mathbf{x}$. For each $i \in [k]$ the family $\mathcal{D}_i$ consists of 2^{k-1} pairs of coordinates corresponding to sets T and $T \triangle \{i\}$. (Here, $\triangle$ denotes the symmetric difference of sets such as $\{1,4,5\} \triangle \{4\} = \{1,5\}$, and $\{1,4,5\} \triangle \{2\} = \{1,2,4,5\}$). Clearly, each family $\mathcal{D}_i$ partitions the universe $[n]$. It is not difficult to verify that if the encoding of $\mathbf{x}$ is corrupted in a δ fraction of coordinates then at least $1 - 2\delta$ fraction of pairs $(T, T \triangle \{i\}) \in \mathcal{D}_i$ fall onto uncorrupted locations. Thus the decoder that XORs the values at these locations obtains the correct value of the i-th bit of $\mathbf{x}$.

The main parameters of interest in locally decodable codes [KT00] are the codeword length n and the query complexity r. In all applications one would ideally like to have both of these parameters as small as possible. One however can not minimize the length and the query complexity simultaneously. There is a trade-off. Determining the true shape of this trade-off is a major open problem.

Locally decodable codes have important applications to data transmission and storage, complexity theory, data structures, derandomization, and theory of fault tolerant computation. However their most prominent applications are in cryptography to the design of private information retrieval schemes.

Private Information Retrieval (PIR) schemes [CGKS98,Yek10] are cryptographic protocols designed to safeguard the privacy of database users. They allow clients to retrieve records from replicated public databases while completely hiding the identity of the retrieved records from database owners. In

such protocols, users query each server holding the database. The protocol ensures that each individual server (by observing only the query it receives) gets no information about the identity of the items of user interest.

There is a strong relation between locally decodable codes and private information retrieval schemes. Below we demonstrate the flavor of this relation, presenting a general procedure that obtains an r-server PIR scheme out of any r-query LDC.

Let C be a locally decodable code encoding k-bit messages to n-bit codewords. At the preprocessing stage servers $\mathcal{S}_1, \ldots, S_r$ encode the k-bit database $\mathbf{x}$ with the code C. Next the user $\mathcal{U}$ who is interested in obtaining the i-th bit of $\mathbf{x}$, tosses random coins and generates an r-tuple of queries $(\mathrm{que}_1, \ldots, \mathrm{que}_r) \in \mathcal{D}_i$. For every $j \in [r]$, the user sends the query que_j to the server $\mathcal{S}_j$. Each server $\mathcal{S}_j$ responds with a one bit answer $C(\mathbf{x})_{\mathrm{que}_j}$. The user combines servers' responses to obtain $\mathbf{x}(i)$.

It is straightforward to verify that the protocol above is private since by the smoothness property for every $j \in [r]$ the query que_j is uniformly distributed over the set of codeword coordinates. The total communication is given by $r(\log n+1)$. Thus short codes of low query complexity yield communication efficient PIR schemes involving a small number of servers.

In this talk we survey the state of the art in locally decodable code and discuss their applications to private information retrieval schemes. We give a high level review of existing main families of codes, focusing on the main ideas behind them. A detailed survey of a large body of work on LDCs (including a detailed treatment of the constructions, lower bounds, and applications) can be found in [Yek12].

References

[CGKS98] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. *Journal of the ACM*, 45:965–981, 1998.

[KT00] Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *32nd ACM Symposium on Theory of Computing (STOC)*, pages 80–86, 2000.

[Yek10] Sergey Yekhanin. Private information retrieval. *Communications of the ACM*, 53(4):68–73, 2010.

[Yek12] Sergey Yekhanin. Locally decodable codes. *Foundations and trends in theoretical computer science*, 2012. To appear. Preliminary version available for download at http://research.microsoft.com/en-us/um/people/yekhanin/Papers/LDC_now.pdf.