Multiset Combinatorial Batch Codes

Hui ZHANG

joint work with

Eitan Yaakobi, Natalia Silberstein
and Srimanta Bhattacharya

Computer Science Department
Technion

January 22, 2017

Table of contents

- A *batch code* is an encoding scheme that encode a large database of n bits to m devices, called *buckets* or *servers*.
- Batch codes can be viewed as several combinatorial objects, including *expanders* and *locally decodable codes*.
- The goal is to minimize the maximal load (number of bits read) on any of the m devices, while also minimizing the total amount of storage (number of bits in the encoded strings) used.

Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Batch codes and their applications," in *Proc. of the 36-sixth Annual ACM Symposium on Theory of Computing STOC '04*, pp. 262–271, 2004.

Definition

An (n, N, k, m, t) *batch code* (BC) over an alphabet Σ, encodes a string $x \in \Sigma^n$ into an m-tuple of strings $y_1, \ldots, y_m \in \Sigma^*$ (called *buckets* or *servers*) of total length N, such that for each k-tuple (called *batch* or *request*) of distinct indices $i_1, \ldots, i_k \in [n]$, the k data items $x_{i_1}, \ldots, x_{i_k}$ can be decoded by reading at most t symbols from each server.

Combinatorial batch codes (CBC), is a special type of batch codes in which all encoded symbols are copies of the input items.

Example

An $(n = 7, N = 15, k = 5, m = 5, t = 1)$-BC

x_1	x_2	x_3	x_4	x_5
x_6	x_6	x_6	x_6	x_6
x_7	x_7	x_7	x_7	x_7

It is also a CBC.


M. B. Paterson, D. R. Stinson, and R. Wei, "Combinatorial batch codes," *Adv. Math. Commun.*, vol. 3, pp. 13–27, 2009.

For the case there are k distinct users wish to directly retrieve data from same devices, the *multiset batch code* is defined.

Definition

An (n, N, k, m, t) *multiset batch code* (MBC) is an (n, N, k, m, t)-BC which also satisfies the following property: For any multiset request of k indices $i_1, \ldots, i_k \in [n]$ there is a partition of the buckets into k subsets $S_1, \ldots, S_k \subseteq [m]$ such that each item $x_{i_j}, j \in [k]$, can be retrieved by reading at most t symbols from each bucket in S_j.

Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Batch codes and their applications," in *Proc. of the 36-sixth Annual ACM Symposium on Theory of Computing STOC '04*, pp. 262–271, 2004.

Example

An $(n = 6, N = 14, k = 4, m = 7, t = 1)$-MBC

x_1	x_2	x_3	$x_1 + x_2$	$x_1 + x_3$	$x_2 + x_3$	$x_1 + x_2 + x_3$
x_4	x_5	x_6	$x_4 + x_5$	$x_4 + x_6$	$x_5 + x_6$	$x_4 + x_5 + x_6$

For example, the request $\{x_1, x_1, x_4, x_4\}$ can be read from buckets $\{\{x_1\}, \{x_2\}, \{x_6\}, \{x_1 + x_2\}, \{x_4 + x_6\}, \{x_5 + x_6\}, \{x_4 + x_5 + x_6\}\}$.

Example

An $(n = 6, N = 14, k = 4, m = 7, t = 1)$-MBC

x_1	x_2	x_3	$x_1 + x_2$	$x_1 + x_3$	$x_2 + x_3$	$x_1 + x_2 + x_3$
x_4	x_5	x_6	$x_4 + x_5$	$x_4 + x_6$	$x_5 + x_6$	$x_4 + x_5 + x_6$

For example, the request $\{x_1, x_1, x_4, x_4\}$ can be read from buckets $\{\{x_1\}, \{x_2\}, \{x_6\}, \{x_1 + x_2\}, \{x_4 + x_6\}, \{x_5 + x_6\}, \{x_4 + x_5 + x_6\}\}$.

- CBCs are not MBC and don't allow to request an item more than once.
- Motivated by the works on codes which enable parallel reads for different users in distributed storage systems, for example, the codes with locality and availability, we introduce a generalization of CBCs, named *multiset combinatorial batch codes* (MCBC).


A. S. Rawat, D. S. Papailiopoulos, A. G. Dimakis, and S. Vishwanath, "Locality and availability in distributed storage," *IEEE Trans. Inform. Theory*, vol. 62, pp. 4481–4493, 2016.


A. Zeh and E. Yaakobi, "Bounds and constructions of codes with multiple localities," in *Proc. IEEE Intl. Symp. Inform. Theory*, Barcelona, Spain, pp. 640–644, Jul. 2016.

Definition

An $(n, N, k, m, t; r)$ *multiset combinatorial batch code (MCBC)* is a collection of subsets of $[n]$, $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ (called *servers*) where $N = \sum_{j=1}^{m} |C_j|$, such that for each multiset request $\{i_1, i_2, \ldots, i_k\}$, in which every element in $[n]$ has multiplicity at most r, there exist subsets $D_1, \ldots, D_m$, where for all $j \in [m]$, $D_j \subseteq C_j$ with $|D_j| \leq t$, and the multiset union[a] of D_j for $j \in [m]$ contains the multiset request $\{i_1, i_2, \ldots, i_k\}$.

[a] For any $i \in [n]$, the multiplicity of i in the multiset union of D_j, $j \in [m]$ is the number of subsets that contain i, that is $|\{j \in [m] : i \in D_j\}|$.

Example

An $(n = 5, N = 15, k = 5, m = 5, t = 1; r = 2)$-MCBC

1	1	2	2	3
3	4	3	4	4
5	5	5	5	5

where the i-th column contains the indices of items stored in a server $C_i \in \mathcal{C}$, $i \in [5]$.

- Our goal is to minimize the total storage N given the parameters n, m, k, t and r of an MCBC.
- Let $N(n, k, m, t; r)$ be the smallest N such that an $(n, N, k, m, t; r)$-MCBC exists.
- An MCBC is called *optimal* if N is minimal given n, m, k, t, r.

When $t = 1$, we omit t from the notation.

- $(n, N, k, m; r)$-MCBC
- $N(n, k, m; r)$

- A *set system* is a pair $(V, \mathcal{C})$, where V is a finite set of *points* and $\mathcal{C}$ is a collection of subsets of V (called *blocks*).
- Given a set system $(V, \mathcal{C})$ with a points set $V = \{v_1, v_2, \ldots, v_n\}$ and a blocks set $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$, its *incidence matrix* is an $m \times n$ matrix M, given by

$$M_{i,j} = \begin{cases} 1 & \text{if } v_j \in C_i, \\ 0 & \text{if } v_j \notin C_i. \end{cases}$$

- If M is the incidence matrix of the set system $(V, \mathcal{C})$, then the set system having incidence matrix $M^\top$ is called the *dual set system* of $(V, \mathcal{C})$.

- For $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ be an $(n, N, k, m; r)$-MCBC, let $V = [n]$, then $(V, \mathcal{C})$ is a set system.
- Let dual set system of $(V, \mathcal{C})$ be $(X, \mathcal{B})$ which is given by:

$$X = [m] \text{ and } \mathcal{B} = \{B_1, B_2, \ldots, B_n\}$$

where for each $i \in [n]$, $B_i \subseteq X$ consists of the indices of servers that store the i-th item.

Example

An $(n = 20, N = 80, k = 16, m = 16)$-CBC $(V, \mathcal{C})$ with $V = [20]$, each column contains the indices of items stored in a server $C_i \in \mathcal{C}$ and also forms a block of the set system[a].

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
5	6	7	8	6	5	8	7	7	8	5	6	8	7	6	5
9	10	11	12	12	11	10	9	10	9	12	11	11	12	9	10
13	14	15	16	15	16	13	14	16	15	14	13	14	13	16	15
17	17	17	17	18	18	18	18	19	19	19	19	20	20	20	20

[a] given in [1] based on an *affine plane* of order 4

N. Silberstein and A. Gál, "Optimal combinatorial batch codes based on block designs," *Des. Codes Cryptogr.*, pp. 1–16, 2014.

Example

Incidence matrix of the $(n = 20, N = 80, k = 16, m = 16)$-CBC $(V, \mathcal{C})$ with 16 rows and 20 columns.

$$\begin{pmatrix}
1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0\\
0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&1&0&0&0\\
0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&1&0&0&0\\
0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&1&0&0&0\\
1&0&0&0&0&1&0&0&0&0&0&1&0&0&1&0&0&1&0&0\\
0&1&0&0&1&0&0&0&0&0&1&0&0&0&0&1&0&1&0&0\\
0&0&1&0&0&0&0&1&0&1&0&0&1&0&0&0&0&1&0&0\\
0&0&0&1&0&0&1&0&1&0&0&0&0&1&0&0&0&1&0&0\\
1&0&0&0&0&0&1&0&0&1&0&0&0&0&0&1&0&0&1&0\\
0&1&0&0&0&0&0&1&1&0&0&0&0&0&1&0&0&0&1&0\\
0&0&1&0&1&0&0&0&0&0&0&1&0&1&0&0&0&0&1&0\\
0&0&0&1&0&1&0&0&0&0&1&0&1&0&0&0&0&0&1&0\\
1&0&0&0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&1\\
0&1&0&0&0&0&1&0&0&0&0&1&1&0&0&0&0&0&0&1\\
0&0&1&0&0&1&0&0&1&0&0&0&0&0&0&1&0&0&0&1\\
0&0&0&1&1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1
\end{pmatrix}$$

Example

Here, the dual set system is $X = [16]$ and $\mathcal{B}$ is as follows.

$$
\begin{array}{cccc}
\{1,5,9,13\} & \{2,6,10,14\} & \{3,7,11,15\} & \{4,8,12,16\} \\
\{1,6,11,16\} & \{2,5,12,15\} & \{3,8,9,14\} & \{4,7,10,13\} \\
\{1,8,10,15\} & \{2,7,9,16\} & \{3,6,12,13\} & \{4,5,11,14\} \\
\{1,7,12,14\} & \{2,8,11,13\} & \{3,5,10,16\} & \{4,6,9,15\} \\
\{1,2,3,4\} & \{5,6,7,8\} & \{9,10,11,12\} & \{13,14,15,16\}
\end{array}
$$

Theorem

The set system $(V, \mathcal{C})$ *is an* (n, N, k, m)*-CBC if and only if its dual set system* $(X, \mathcal{B})$ *satisfies the following* Hall's condition*:*

for all $h \in [k]$*, and any* h *distinct blocks* $B_{i_1}, B_{i_2}, \ldots, B_{i_h} \in \mathcal{B}$, $|\cup_{j=1}^{h} B_{i_j}| \geq h$.

M. B. Paterson, D. R. Stinson, and R. Wei, "Combinatorial batch codes," *Adv. Math. Commun.*, vol. 3, pp. 13–27, 2009.

Theorem

The set system $(V, \mathcal{C})$ *is an* $(n, N, k, m; r)$*-MCBC if and only if its dual set system* $(X, \mathcal{B})$ *satisfies the following* multiset Hall's condition*:*

for all $h \in \left[\left\lceil \frac{k}{r} \right\rceil\right]$*, and any* h *distinct blocks* $B_{i_1}, B_{i_2}, \ldots, B_{i_h} \in \mathcal{B}$*,* $|\cup_{j=1}^{h} B_{i_j}| \geq \min\{hr, k\}$.

Theorem

1. $N(n,k,m;r) \geq rn$.
2. $N(n,k,m;r) \geq N(n,k,m;i)$ *for* $i \in [r-1]$.
3. $\frac{1}{r}N(rn,k,m) \leq N(n,k,m;r) \leq N(rn,k,m)$.
4. $N(n,k,m;r) \leq rN(n,\left\lceil \frac{k}{r} \right\rceil, \left\lfloor \frac{m}{r} \right\rfloor)$.
5. $N(n,k,m;k) = nk$.

A *Steiner system* $S(2, \ell, m)$ is a set system $(X, \mathcal{B})$, where X is a set of m points, $\mathcal{B}$ is a collection of ℓ-subsets (blocks) of X, such that each pair of points in X occurs together in exactly one block of $\mathcal{B}$.

Theorem

Let $(X, \mathcal{B})$ be an $S(2, \ell, m)$ with b blocks and $m > \ell$. Then the dual set system of $(X, \mathcal{B})$ is a $(b, b\ell, k, m; r)$-MCBC for any $\lfloor \frac{\ell}{2} \rfloor + 1 \leq r \leq \ell$ and $k \leq (\ell - r + 1)(2r - 1)$.

We only need to check the largest k, i.e., $k = (\ell - r + 1)(2r - 1)$.

Proof.

- If $h \leq \ell + 1 \leq |\mathcal{B}|$, the union of any h blocks contains at least $\ell + (\ell - 1) + \cdots + (\ell - (h-1)) = h\ell - \binom{h}{2}$ points.
- Let us consider some h blocks $B_{i_1}, B_{i_2}, \ldots, B_{i_h}$ with $1 \leq h \leq \left\lceil \frac{k}{r} \right\rceil = \left\lceil \frac{(\ell-r+1)(2r-1)}{r} \right\rceil \leq 2(\ell - r) + 2$.
- If $h \in [2(\ell - r) + 1]$, then $r \leq \ell - \frac{h-1}{2}$, and $|\cup_{j=1}^{h} B_{i_j}| \geq h\ell - \binom{h}{2} = h\left(\ell - \frac{h-1}{2}\right) \geq hr = \min\{hr, k\}$.
- If $h = 2(\ell - r) + 2$, then $|\cup_{j=1}^{h} B_{i_j}| \geq \ell h - \binom{h}{2} = (\ell - r + 1)(2r - 1) = k$.
- Therefore, the multiset Hall's condition holds for any $1 \leq h \leq \left\lceil \frac{k}{r} \right\rceil$.

Taking Affine Plane $S(q^2, q, 1)$, we get:

Corollary

There exists a $(q^2+q, q^3+q^2, k, q^2; r)$*-MCBC for any* $\lfloor \frac{q}{2} \rfloor + 1 \leq r \leq q$ *and* $k \leq (q-r+1)(2r-1)$ *for any prime power* q*.*

Example

An $S(2,4,16)^a$ $(X,\mathcal{B})$ with $X=[16]$ and $\mathcal{B}$ is as follows.

$$\begin{array}{cccc}
\{1,5,9,13\} & \{2,6,10,14\} & \{3,7,11,15\} & \{4,8,12,16\} \\
\{1,6,11,16\} & \{2,5,12,15\} & \{3,8,9,14\} & \{4,7,10,13\} \\
\{1,8,10,15\} & \{2,7,9,16\} & \{3,6,12,13\} & \{4,5,11,14\} \\
\{1,7,12,14\} & \{2,8,11,13\} & \{3,5,10,16\} & \{4,6,9,15\} \\
\{1,2,3,4\} & \{5,6,7,8\} & \{9,10,11,12\} & \{13,14,15,16\}
\end{array}$$

[a] an *affine plane* of order 4

N. Silberstein and A. Gál, "Optimal combinatorial batch codes based on block designs," *Des. Codes Cryptogr.*, pp. 1–16, 2014.

Example

An $(n = 20, N = 80, k, m = 16; r)$-MCBC $(V, \mathcal{C})$ with $V = [20]$.

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
5	6	7	8	6	5	8	7	7	8	5	6	8	7	6	5
9	10	11	12	12	11	10	9	10	9	12	11	11	12	9	10
13	14	15	16	15	16	13	14	16	15	14	13	14	13	16	15
17	17	17	17	18	18	18	18	19	19	19	19	20	20	20	20

It is possible to verify that it gives a $(20, 80, k, 16; r)$-MCBC for $(k, r) \in \{(16, 1), (11, 2), (10, 3), (7, 4)\}$.

- A regular $(n, N, k, m; r)$-MCBC is an MCBC in which each server stores the same number μ of items, where $\mu = N/m$.

- Given n, m, k, r, let $\mu(n, k, m; r)$ denote the smallest number of items stored in each server, then the optimal value of N is determined by $\mu(n, k, m; r)$, that is, $N = m\mu(n, k, m; r)$.

Lemma

$$\mu(n,k,m;r) \geq \left\lceil \frac{N(n,k,m;r)}{m} \right\rceil.$$

Proof.

Since a regular $(n,N,k,m;r)$-MCBC is also an $(n,N,k,m;r)$-MCBC, then $m\mu(n,k,m;r) \geq N(n,k,m;r)$. □

If $r = k$, then $\mu(n, k, m; k) \geq kn/m$.

Theorem

$\mu(n, k, m; k) = \frac{kn}{m}$ if and only if $n = c \cdot \frac{m}{\gcd(m,k)}$ for some integer $c \geq 0$.

Let's recall the definition of multiset batch codes.

Definition

An (n, N, k, m, t) *multiset batch code* (MBC) is an (n, N, k, m, t) batch code which also satisfies the following property: For any multiset request of k indices $i_1, \ldots, i_k \in [n]$ there is a partition of the buckets into k subsets $S_1, \ldots, S_k \subseteq [m]$ such that each item $x_{i_j}, j \in [k]$, can be retrieved by reading at most t symbols from each bucket in S_j.

Let $N^B(n, k, m)$ denote the minimum N for any $(n, N, k, m, 1)$ multiset batch code.

H. Zhang, S. Bhattacharya, E. Yaakobi and N. Silberstein, "On Two-Dimensional Codes with Availability: Bounds and Constructions".

Example

An $(n = 6, N = 14, k = 4, m = 7, t = 1)$-MBC

x_1	x_2	x_3	$x_1 + x_2$	$x_1 + x_3$	$x_2 + x_3$	$x_1 + x_2 + x_3$
x_4	x_5	x_6	$x_4 + x_5$	$x_4 + x_6$	$x_5 + x_6$	$x_4 + x_5 + x_6$

For example, the request $\{x_1, x_1, x_4, x_4\}$ can be read from buckets $\{\{x_1\}, \{x_2\}, \{x_6\}, \{x_1 + x_2\}, \{x_4 + x_6\}, \{x_5 + x_6\}, \{x_4 + x_5 + x_6\}\}$.

$\Rightarrow$ By Gadget Lemma

Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Batch codes and their applications," in *Proc. of the 36-sixth Annual ACM Symposium on Theory of Computing STOC '04*, pp. 262–271, 2004.

Theorem

$N^B(n, k, k+1) \geq n(k-1) + \lceil \frac{n}{2} \rceil$.

Proof.

- Given an $(n, N, k, m = k+1, 1)$ MBC, it may be observed that each of the n information symbols appears at least $k-1$ times systematically in the encoding.
- We delete $k-2$ of these $k-1$ singletons for each information symbol.
- Now, the resulting code $\mathcal{C}$ is systematic with dimension n, distance at least 2, and locality at most 2.
- For $\mathcal{C}$, we have $N \geq n + \lceil \frac{n}{2} \rceil$. Hence, $N^B(n, k, k+1) \geq n(k-1) + \lceil \frac{n}{2} \rceil$.

Theorem

$$N^B(n,k,m) \leq n(k-1) + \left\lceil \frac{n}{m-k+1} \right\rceil.$$

Proof.

- $N^B(n,k,m) \leq N^B(n,k-2,k-2) + N^B(n,2,m-k+2)$.
- $N^B(n,k-2,k-2) = n(k-2)$. (The code with each column contains all the information symbols.) $\Rightarrow \mathcal{C}_1$
- $N^B(n,2,m-k+2) \leq n + \lceil \frac{n}{m-k+1} \rceil$.

  (Consider the $(\lceil \frac{n}{m-k+1} \rceil \times (m-k+2), n, 2, 1)_B$-code, where we partition the n information bits into $\lceil \frac{n}{m-k+1} \rceil$ groups with each group having at most $m-k+1$ bits.) $\Rightarrow \mathcal{C}_2$
- $\mathcal{C} = (\mathcal{C}_1 | \mathcal{C}_2)$.

Corollary

- $N^B(n, k, k+1) = n(k-1) + \lceil \frac{n}{2} \rceil$.
- $N^B(n, 2, m) = n + \lceil \frac{n}{m-1} \rceil$.

Thanks!