Combinatorial Batch Codes

Anna Gál

UT Austin

Joint work with: Natalia Silberstein

Batch codes

[Ishai, Kushilevitz, Ostrovsky, Sahai, 2004]

string x of length $n \rightarrow m$ strings (buckets) such that ANY subset of k symbols from x can be retrieved by reading at most t symbols from each bucket.

Goal: keep t small (e.g. $t = 1$)
minimizing m AND total storage size.

Motivation
Load balancing in distributed storage

Given data set of n items, use m servers for storage

Load of server: number of symbols read from it

Minimize load of servers,

number of servers,

while also minimizing total storage space.

Motivation
Private Information Retrieval (PIR)

DATA: n-bit string x

USER: wishes to retrieve x_i and keep i private

Download entire x: n bits communication

With one server, improvement only possible under computational hardness assumptions [CGKS95]

e.g. $O(n^{\epsilon}), O(\log n)$ bits communication [KO97]

x is held by several servers

User gets i-th item, servers learn nothing about i.

2 servers: $O(n^{1/3})$ [CGKS95]

s servers: $n^{1/\Omega(s)}$ [CGKS95, Amb97, BIKR02]

$\log n$ servers polylog(n) communication

Time complexity of servers remains $\Omega(n)$.

PIR protocol to retrieve 1 bit of an n-bit string

C(n): communication (number of bits transmitted)

T(n): time complexity of servers

What is the cost to retrieve k bits?

Trivial: $\leq kC(n)$ communication, $\leq kT(n)$ time.

Suppose we have a batch code:

n bit string $\rightarrow$ m strings of lengths $N_1, \ldots, N_m$.

$t = 1$: any k bits can be retrieved by reading at most 1 bit from each server

gives k out of n PIR protocol with

$\leq \sum_{i=1}^{m} C(N_i)$ communication

$\leq \sum_{i=1}^{m} T(N_i)$ time

Examples

$m = 3$ servers; repeat x 3 times

To retrieve k bits, read $k/3$ bits/server, $N = 3n$

Can we have storage $N = 1.5n$, and load $< k$?
not possible with just replication for $m = 3$.
($\exists n/2$ bits at one server, $n/6$ bits at same server)

$m = 3$, $N = 1.5n$, load $k/2$
split x in half: $x = (L, R)$, store $L, R, L \oplus R$
(retrieve any 2 bits reading 1 bit/server)

Combinatorial Batch Codes

Name by [Paterson, Stinson, Wei 2008].

Replication only batch codes

Each server gets a subset of the bits of x

Notation: $t = 1$, $(n, N, k, m) - CBC$:

$x \in \{0, 1\}^n \to m$ servers

ANY k bits of x can be retrieved by reading at most 1 bit from each server

N: total storage used

Matrix view

Rows: servers, columns: items

1	0	0	1	1	0
0	1	0	1	0	1
0	0	1	0	1	1

$(n, N, k, m) - CBC$:
Any k columns contain a “diagonal” of size k.

Set system view

$\mathcal{F} = F_1, \ldots, F_m$, where $F_i \subseteq [n]$

F_i specifies which bits stored at server i

$(n, N, k, m) - CBC$:
Any $A \subseteq [n]$ with $|A| = k$ forms a system of distinct representatives for some k members of $\mathcal{F}$.

Graph view

Bipartite graph $G = (V_1, V_2, E)$

$|V_1| = m$ servers, $|V_2| = n$ bits

edge $(i, j) \in E$ if j-th bit is stored at server i

$(n, N, k, m) - CBC$:
n by m bipartite graph, s.t. for any $A \subseteq V_2$ with $|A| = k$
there is a matching of A into some subset of V_1

Hall's Condition

For $A \subseteq V_2$ there is a matching of A into V_1

if and only if

$\forall S \subseteq A$ has $\geq |S|$ neighbours.

$(n, N, k, m) - CBC$:

$\forall S \subseteq V_2$ with $|S| \leq k$ has $\geq |S|$ neighbours.

Bipartite Expander graphs

$G = (V_1, V_2, E)$ is a (k, a)-vertex expander if $\forall S \subseteq V_2$ with $|S| \leq k$ has $\geq a|S|$ neighbours.

$(n, N, k, m) - CBC$: $(k, 1)$-expander.

want to minimize N (number of edges)

Two Trivial CBCs

1. $C(x) = x, x, \ldots, x$; $m = k$, but storage $N = kn$ very large

1	1	1	1	1	1
1	1	1	1	1	1
1	1	1	1	1	1

2. $C(x) = x_1, x_2, \ldots, x_n$ $N = n$, but $m = n$ very large

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Note: $k \leq m \leq n$

(n, N, k, m)-CBC is called OPTIMAL if total storage N is minimal for given n, m and k.

1	0	0	1	1	1	1
0	1	0	1	1	1	1
0	0	1	1	1	1	1

optimal when $m = k$

$N(n, m, k)$ minimal value of N for given n, m, k.

Easy: $n \leq N(n, m, k) \leq kn - m(k - 1)$

Precise values known for $k = 2, 3, 4$ for any n and m; for $m = n, n - 1, n - 2$ for any k, and for $n \geq \binom{m}{k-2}$.

Known bounds [PSW09, BRR12]

n	$N(n,k,m)$
$n \geq (k-1)\binom{m}{k-1}$	$kn - (k-1)\binom{m}{k-1}$
$\binom{m}{k-2} \leq n \leq (k-1)\binom{m}{k-1}$	$n(k-1) - \left\lfloor \frac{(k-1)\binom{m}{k-1}-n}{m-k+1} \right\rfloor$

Last bound generalized [BRR12]
Let $1 \leq c \leq k-1$ be the least integer such that

$$n \leq \frac{(k-1)\binom{m}{c}}{\binom{k-1}{c}}.$$

Then $N(n,k,m) \geq nc - \left\lfloor \frac{(k-c)\left(\frac{(k-1)\binom{m}{c}}{\binom{k-1}{c}}-n\right)}{m-k+1} \right\rfloor \geq n(c-1)$

[BRR12]: Tight for half of the values of n in the range $\binom{m}{k-2} - (m-k+1)A(m,4,k-3) \le n \le \binom{m}{k-2}$

Open if tight, even up to constant factors,

for $n < \binom{m}{k-2} - (m-k+1)A(m,4,k-3)$

$A(m,4,k-3)$: max # of codewords in a binary constant weight code (length m, weight $k-3$ Hamming distance 4)

We construct optimal CBCs for n in this range

Block Designs

Subsets of "points" called "blocks"

1. each block contains exactly ℓ points

2. each pair of points is in exactly λ blocks

Transversal Designs

$TD(\ell, h)$: ℓ groups of points each of size h

1. each block contains one point from each group
2. any pair of points from different groups in 1 block

ℓh points
number of blocks is h^2
number of blocks that contain a given point is h

Resolvable Transversal Designs

$TD(\ell, h)$ **is resolvable** if

set of h^2 blocks can be partitioned into

h classes of h blocks, s.t.

each point is in exactly one block of each class

q prime power,

there exists resolvable $TD(\ell, q)$ for any $\ell \leq q$.

$TD(3,4)$

$$\left(\begin{array}{cccc|cccc|cccc|cccc}
1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0\\
0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0\\
0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0\\
0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1\\
\hline
1&0&0&0&0&1&0&0&0&0&0&1&0&0&1&0\\
0&1&0&0&1&0&0&0&0&0&1&0&0&0&0&1\\
0&0&1&0&0&0&0&1&0&1&0&0&1&0&0&0\\
0&0&0&1&0&0&1&0&1&0&0&0&0&1&0&0\\
\hline
1&0&0&0&0&0&1&0&0&1&0&0&0&0&0&1\\
0&1&0&0&0&0&0&1&1&0&0&0&0&0&1&0\\
0&0&1&0&1&0&0&0&0&0&0&1&0&1&0&0\\
0&0&0&1&0&1&0&0&0&0&1&0&1&0&0&0
\end{array}\right)$$

rows: points, columns: blocks

Optimal CBCs from transversal designs

We construct (n, N, k, m)-CBC for

$n = q^2 + q - 1$, $N = q^3 - 1$, $k = m - 1$, $m = q(q - 1)$.

Construction:
Add incidence vectors of groups to $TD(q - 1, q)$

Optimal CBC

$$
\left(\begin{array}{cccc|cccc|cccc|cccc|ccc}
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
\hline
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
\hline
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right)
$$

Proof ideas

Optimal construction for $m = n$:

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Transversal designs seem to have the right structure for CBCs.

Proof ideas

We have to show,

any set of $r \leq k$ columns (blocks)

covers at least r points

Permutation matrices $\rightarrow$

condition holds for any subset within classes

Resolvability $\rightarrow$ full class covers all m points

Proof ideas

What if we take one column from each class?
(ignore "special" class for now)

that is at most q columns ($r \leq q$)
any one column covers $q - 1$ points,
so we just need one extra point

Use property of TD: any pair of columns
from different classes intersect in one point

Proof sketch

So far, proved condition for any set of $r \leq q$ columns.

Now let $r = iq + j$ $(1 \leq j \leq q)$.
on average, about i columns/class

1. If there is a class with $> i + 1$ columns used
cover enough points by just this class: $(i+2)(q-1) \geq r$

2. Otherwise, largest class covers $(i+1)(q-1)$,
show other classes cover enough additional points
(usually one more class is sufficient)

Affine Plane of order q

exists when q is prime power

q^2 points
$q(q+1)$ blocks of size q
each pair of points in exactly one block

Every affine plane is resolvable:
blocks can be partitioned into $q+1$ classes
(q blocks in each class) s.t.
each point is in exactly one block of each class

"Parallel classes": blocks within a class are disjoint
parallel "lines" (they don't intersect)

Affine plane of order q

$$
\left(\begin{array}{cccc|cccc|cccc|cccc|cccc}
1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0\\
0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&1&0&0&0\\
0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&1&0&0&0\\
0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&1&0&0&0\\
\hline
1&0&0&0&0&1&0&0&0&0&0&1&0&0&1&0&0&1&0&0\\
0&1&0&0&1&0&0&0&0&0&1&0&0&0&0&1&0&1&0&0\\
0&0&1&0&0&0&0&1&0&1&0&0&1&0&0&0&0&1&0&0\\
0&0&0&1&0&0&1&0&1&0&0&0&0&1&0&0&0&1&0&0\\
\hline
1&0&0&0&0&0&1&0&0&1&0&0&0&0&0&1&0&0&1&0\\
0&1&0&0&0&0&0&1&1&0&0&0&0&0&1&0&0&0&1&0\\
0&0&1&0&1&0&0&0&0&0&0&1&0&1&0&0&0&0&1&0\\
0&0&0&1&0&1&0&0&0&0&1&0&1&0&0&0&0&0&1&0\\
\hline
1&0&0&0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&1\\
0&1&0&0&0&0&1&0&0&0&0&1&1&0&0&0&0&0&0&1\\
0&0&1&0&0&1&0&0&1&0&0&0&0&0&0&1&0&0&0&1\\
0&0&0&1&1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1
\end{array}\right)
$$

Uniform CBC

Each item stored at same number of servers.

Graph view: d-regular bipartite expander

Probabilistic constructions known - not optimal.

Optimal constructions were known for $d = 2, k-1, k-2$

We give optimal constructions for $d = \sqrt{k}$.

Affine plane is uniform CBC, with $k = m = q^2$

OPEN PROBLEMS

Is the bound

$$N(n,k,m) \geq nc - \left\lfloor \frac{(k-c)\left(\frac{(k-1)\binom{m}{c}}{\binom{k-1}{c}} - n\right)}{m-k+1} \right\rfloor$$

always tight?

Optimal uniform CBCs for other values of d