

# Efficient Stream Computation of Approximate Wavelet Representations

Anna C. Gilbert  
Yannis Kotidis  
S. Muthukrishnan  
Martin J. Strauss

AT&T Labs—Research

## Streaming Model (1 Terabyte/day) —

AT&T's network, daily traffic:

- 300M calls
- 1 Terabyte of billing data, generated one call at a time

$a_i = \#$  calls from phone  $i$ ,  $0 \leq i < N = 80M$ .

Suppose  $a$  is compactly representable.

How to process  $a$  into a compact representation?

## More Streaming Examples \_\_\_\_\_

Each Walmart cash register emits a *stream*:

*6 beers, 24 beers, 40 diapers, 6 beers, ...*

Defines *signal*  $a$ :  $a_i = \#$  items of type  $i$ , e.g.,

$$a_{\text{beer}} = 36, a_{\text{diapers}} = 40, \dots$$

Want good and compact representation for  $a$  without ...

- storing all of  $a$
- spending too much time processing items
- spending too much time post-processing

## Good Haar Wavelet Representations \_

Often, a few wavelet (or w. packet) coefficients capture significant *energy* fraction  $\eta$ :

$$\left\| a - \sum_{k=1}^B d_{j_k} \psi_{j_k} \right\|_2^2 \leq (1 - \eta) \|a\|_2^2,$$

where  $B/\eta \ll N$  is small. (E.g.,  $B = 10$  terms capture  $\eta = 20\%$  of the energy.)

We *assume* good representations exist.

Problem: How to find good rep'ns from the stream (or report that there are none)?

## Naive Schemes

---

1. Traditional: Each item type gets counter,  $a_i$ . At end,  $O(N)$ -time wavelet decomposition. But...
  - high storage and post-processing cost
2. Find best representation from a sample or prefix of data. Update these coefficients from the stream. But...
  - Scheme fails if data changes character after sample.
  - Hard to sample from stream of canceling positive & negative items.

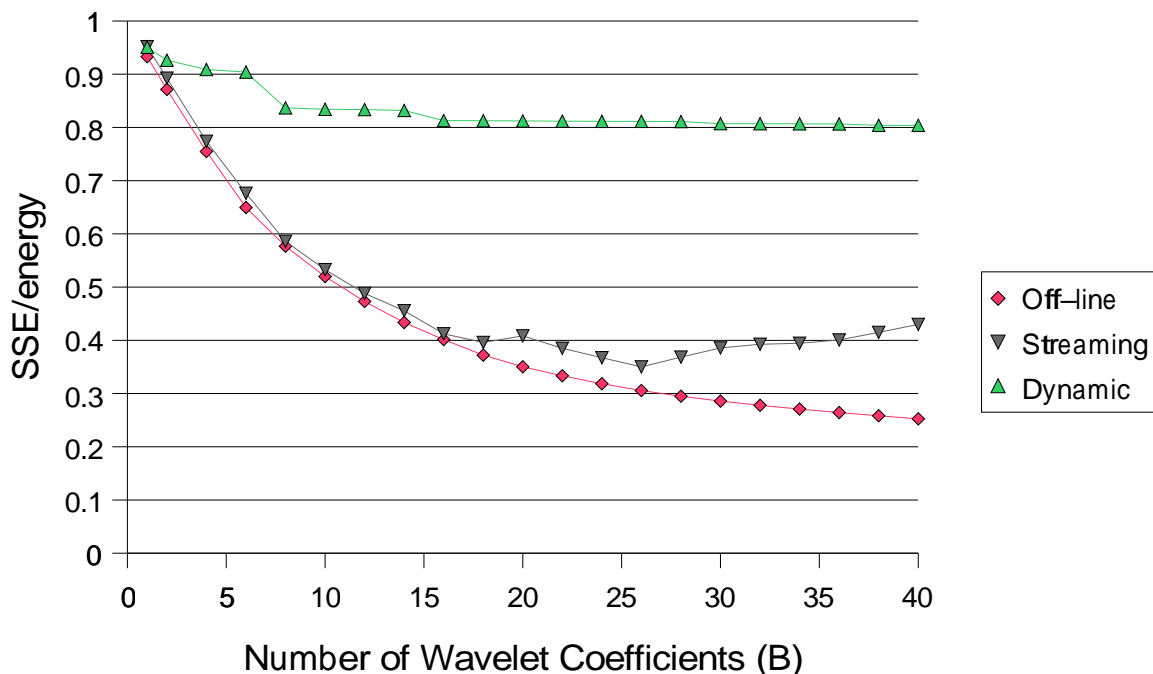
## Example—AT&T Data

---

$a_i$  is the number of telephone calls made from area code and exchange  $i$ , over a week.

$N \approx 60k$ , streaming space 4k, in words, for estimation only (Naive post-processing procedure.)

Our quality is much better than sticking with the best coefficients for Monday's data.



## Space Savings

---

Traditional: Each item type gets counter,  $a_i$ .  
At end,  $O(N)$ -time wavelet decomposition.

- Storage  $O(N)$  (*bad*)
- Post-processing time  $O(N)$  (*bad*)
- +  $O(1)$  time per item (*good*)

## Space Savings

---

Traditional: Each item type gets counter,  $a_i$ .  
At end,  $O(N)$ -time wavelet decomposition.

- Storage  $O(N)$  (*bad*)
- Post-processing time  $O(N)$  (*bad*)
- +  $O(1)$  time per item (*good*)

Us: Make small randomized *sketch*  $h(a)$  of  $a$  from stream. Reconstruct  $a$  from sketch via  $\rho$ .

- + Storage  $\tilde{O}(1)$  (*good*)
- + Post-processing time  $\tilde{O}(1)$  (*good*)
  - Error increases, by  $\epsilon$  (*as little as desired*)
  - Probability  $\delta$  of failure (*as little as desired*)
  - Seek  $B$  terms (*as many as desired*)
  - Assume capture  $\eta \|a\|_2^2$  (*as little as desired*)
- Per-item time  $\tilde{O}(1)$  (*bounded loss*)

Here

$$\tilde{O}(f(N)) = f(N)(\log(N)B\eta^{-1}\epsilon^{-1}\log(1/\delta))^{O(1)}$$

highlights significant dependence on  $N$ .

## Data Input Models \_\_\_\_\_

Traditional (*ordered aggregate* model): Get  $a_0, a_1, a_2, \dots$ , in order

Unordered aggregate model: Get  $\{a_0, a_1, a_2 \dots\}$  in *unknown* order

Cash-register model:  $a_i$  implied by stream of transactions. (Hardest model to process.)

## Cancelation Possible in Cash Register Data \_\_\_\_\_

Data items may be positive or negative, and can cancel. E.g.,

*6 beers, 24 beers, 2 diapers, return 30 beers,  
...*

Good representation: 2 diapers.

We succeed despite canceling beers (sampling from stream won't work).

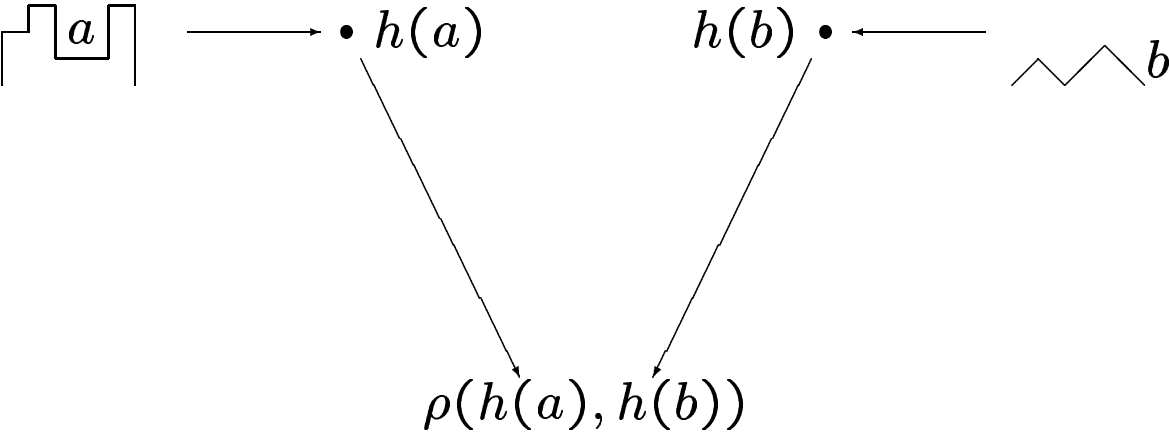
# Low Communication, Example \_\_\_\_\_

Walmart-Minneapolis sells vector  $a$ ; St. Paul sells  $b$ . Want representation of  $a + b$ ...

- without too much total communication.

Our solution:

Use common sketching function  $h$ . Exchange small sketches  $h(a)$  &  $h(b)$ , whence  $\approx a + b$ .



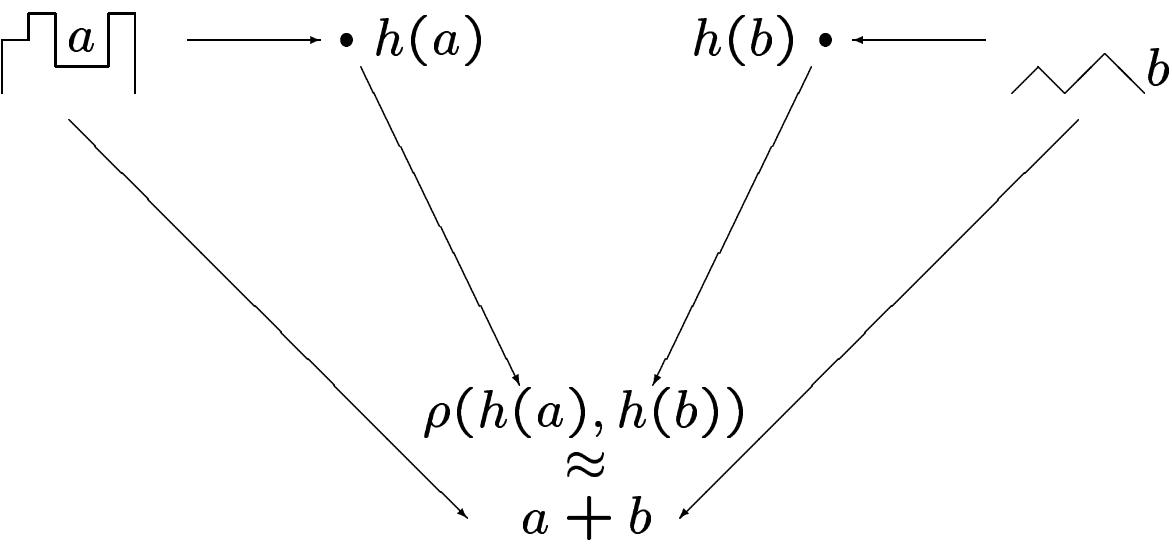
# Low Communication, Example \_\_\_\_\_

Walmart-Minneapolis sells vector  $a$ ; St. Paul sells  $b$ . Want representation of  $a + b$ ...

- without too much total communication.

Our solution:

Use common sketching function  $h$ . Exchange small sketches  $h(a)$  &  $h(b)$ , whence  $\approx a + b$ .



## Periodic Data, Example \_\_\_\_\_

Minneapolis gets a new vector  $a^{(t)}$  on each day. At the end of the month, they want an approximate representation about a linear combination  $\sum w_t a^{(t)}$  of sales without

- predicting  $w_t$
- storing each day's data.

Store just a small *sketch* of each days' data. From the sketches, get an approximation of the linear combination of sales.

## Stream Models Disk/Tape Storage —

Massive data stored on disk or tape.

Data Analyst with small-memory PC on shared file system.

Streaming disk/tape access much faster than random access.

## Approximate Representations—Uses \_

We want an approximate representation with guarantees. Useful

- directly, (e.g., for noisy data)
- for exploratory data analysis (approximate is cheaper/faster than exact)
- in database query optimization: How best to retrieve records of

employees of age 29–34 in MN -?

Depends on approximate number of employees age 29–34 and number of employees in MN.

## Easy and Hard Cases \_\_\_\_\_

EASY: Computing ordinary wavelets from ordered aggregate data is easy—work on one coefficient at a time per resolution level.

E.g., while reading  $a_2 \dots$

- done with  $a_1 - a_0$
- working on  $a_3 - a_2$
- not yet working on  $a_5 - a_4$ .

Store only largest coefficients.

HARD: Wavelet packets are hard even in ordered aggregate model.

HARD: Ordinary wavelets are hard in unordered aggregate model. (Note: Aggregate data feeds via packet networks are generally unordered.)

HARDEST: Cash-register model has updates and cancelation, and is harder still.

## Main Theorem ---

We give sketch function  $h$  and reconstruction  $\rho$  such that, if

$$\left\| a - \sum_{k=1}^B d_{j_k} \psi_{j_k} \right\|_2^2 \leq (1 - \eta) \|a\|_2^2,$$

then, except with probability  $\delta$ ,

$$\|a - \rho(h(a))\|_2^2 \leq (1 - \eta + \epsilon\eta) \|a\|_2^2.$$

Furthermore,

- the space used by  $h$ ,
- the per-item time used by  $h$ , and
- the post-processing time used by  $\rho$

is each a power of  $(\log(N)B \log(1/\delta)\eta^{-1}\epsilon^{-1})$ .

## Main Theorem

---

We give sketch function  $h$  and reconstruction  $\rho$  such that, if

$$\left\| a - \sum_{k=1}^B d_{j_k} \psi_{j_k} \right\|_2^2 \leq (1 - \eta) \|a\|_2^2,$$

then, except with probability  $\delta$ ,

$$\|a - \rho(h(a))\|_2^2 \leq (1 - \eta + \epsilon\eta) \|a\|_2^2.$$

Furthermore,

- the space used by  $h$ ,
- the per-item time used by  $h$ , and
- the post-processing time used by  $\rho$

is each a power of  $(\log(N)B \log(1/\delta)\eta^{-1}\epsilon^{-1})$ .

Notes:

- $\rho(h(a))$  is a  $B$ -term representation.
- $h$  is randomized;  $\rho$  is deterministic.
- $a$  in cash-register or aggregate format.
- if  $a$  has no  $B$ -term  $(\eta - \epsilon\eta) \|a\|_2^2$ -energy rep'n, the algorithm reports “no  $\eta$  rep'n.” (Either behavior if  $(\eta - \epsilon\eta) \|a\|_2^2 < \text{energy} < \eta \|a\|_2^2$ .)

## Weakest Possible Assumptions \_\_\_\_\_

We assume some  $B$ -term approximation over wavelets  $\psi_j$  captures an  $\eta$  fraction of the energy:

$$\left\| a - \sum_{k=1}^B d_{j_k} \psi_{j_k} \right\|_2^2 \leq (1 - \eta) \|a\|_2^2,$$

where  $B/\eta \ll N$  for good performance.

We *do not* assume:

- smoothness of the (discrete!) data
- a probabilistic model for the data source
- $a$  is a superposition  $a = \sum_{k=1}^B d_{j_k} \psi_{j_k}$

▷ Weakest possible assumption for our results.

Unknown  $B$  and  $\eta$  \_\_\_\_\_

What if we don't know  $B$  and  $\eta$  in advance?

We find each coefficient  $d$  with  $d^2 \geq \frac{1}{M} \|a\|_2^2$ , where  $M$  is the available computational resources.

We try *all*  $B$  and smallest possible  $\eta$ , where  $B/\eta \leq M$ .

Thus, for each  $B \leq M$ , we find a  $B$ -term representation with at least  $B/M$  of the energy, if a slightly better one exists.

## Sources of Error

---

Our representation's error has 4 sources:

- $N - B$  dropped wavelet terms
- Dropped (small) terms among the  $B$
- Our coeff of  $\psi$  only approximates  $\langle a, \psi \rangle$
- We can't rank close coeffs—if  $(B + 1)$  coeffs are about equally good, we find all  $B + 1$  and return arbitrary  $B$  of them.

We bound all sources of error.

▷ Traditional approximations only err from the  $N - B$  dropped terms.

By allowing other *sources* of error and tiny probability of failure, we achieve better computational cost with *magnitude* of error as close as desired to best possible offline rep'n.

## Lower bounds

---

An algorithm that finds the best  $B$ -term representation on unordered aggregate data

- without degrading by  $(1 + \epsilon)$ , *or*
- if the best rep'n is not very good, *or*
- with zero probability of failure

provably uses at least  $N / \log^{O(1)}(N)$  streaming space—almost enough to store the signal.

Thus we *must*

- approximate the best  $B$ -term rep'n, *and*
- assume that there is a good rep'n, *and*
- use a randomized algorithm.

Issues are algorithmic, not analytic/geometric.

## Main Theorem ---

We give sketch function  $h$  and reconstruction  $\rho$  such that, if

$$\left\| a - \sum_{k=1}^B d_{j_k} \psi_{j_k} \right\|_2^2 \leq (1 - \eta) \|a\|_2^2,$$

then, except with probability  $\delta$ ,

$$\|a - \rho(h(a))\|_2^2 \leq (1 - \eta + \epsilon\eta) \|a\|_2^2.$$

Furthermore,

- the space used by  $h$ ,
- the per-item time used by  $h$ , and
- the post-processing time used by  $\rho$

is each a power of  $(\log(N)B \log(1/\delta)\eta^{-1}\epsilon^{-1})$ .

## Algorithm Overview \_\_\_\_\_

Two phases:

- Identify large coefficients
- Estimate their values

(Finding large coefficients suffices for a good representation.)

Both phases use a streaming algorithm for  $\ell^2$  norms.

## Estimating $\|\cdot\|_2^2$

---

From Alon-Matias-Szegedy.

There are sketching function  $h'$  and reconstruction  $\rho'$  such that, except with probability  $\delta$ ,

$$\rho'(h'(a)) = (1 \pm \epsilon)\|a\|_2^2.$$

- The space used by  $h'$ ,
- The per-item time used by  $h'$ , and
- The time used by  $\rho'$

is each  $\log^{O(1)}(N) \log(1/\delta)/\epsilon^2$ .

Henceforth, we call  $h'$  an “ $\ell^2$  sketch” of  $a$ .

Estimating  $\|\cdot\|_2^2$ , algorithm \_\_\_\_\_

Idea: Pseudorandom Johnson-Lindenstrauss

Choose multiple repeated copies of  $r_i = \pm 1$  uniformly at random.

Sketch:  $h'(a)$  is  $\log(1/\delta)/\epsilon^2$  repeated copies of  $\langle a, r \rangle$ , a random projection.

Reconstruct:

$$\rho'(h'(a)) = \text{median}_{\log(1/\delta)} \text{mean}_{1/\epsilon^2}(h'(a))^2,$$

the median of  $\log(1/\delta)$  copies of averages of  $1/\epsilon^2$  copies of  $h'(a)$ .

Estimating  $\|\cdot\|_2^2$ , proof \_\_\_\_\_

Put  $X = (\sum_i a_i r_i)^2$ .

$$\begin{aligned} E[X] &= E \left[ \sum_i a_i^2 r_i^2 + \sum_{i \neq j} a_i a_j r_i r_j \right] \\ &= \sum_i a_i^2, \end{aligned}$$

since  $r_i^2 \equiv 1$ ,  $E[r_i] = E[r_j] = 0$  and  $r_i, r_j$  are independent. Similarly,

$$E[X^2] \leq O(E^2[X]),$$

using independence of any four  $r_i$ 's. Take median of means to improve distortion  $\epsilon$  and failure probability  $\delta$ , using Chebychev and Chernoff (standard).

Need only 4-wise independent  $r_i$ 's. From random seed  $s$ ,  $|s| \leq \log^2(N)$ , construct  $N$   $r_i$ 's as needed. (Saves space.)

## Coefficient Estimation from Distances

Use the Lemma to estimate  $\langle a, \psi \rangle$  to within  $\pm \epsilon \|a\|_2$  additively:

Let  $u_a = a/\|a\|_2$ . (Note  $\|\psi\|_2 = 1$ .)

$$\langle a, \psi \rangle = \|a\|_2 \left(1 - \|u_a - \psi\|_2^2/2\right).$$

Estimate  $\|u_a - \psi\|_2^2$  to within

$$(\epsilon/4)\|u_a - \psi\|_2^2 \leq \epsilon.$$

Get  $\langle a, \psi \rangle$  to within  $\epsilon \|a\|_2$ , additively.

Use the same linear  $h'$  as for  $\ell^2$ ; reconstruction is different.

## Estimate each Wavelet Coefficient \_\_\_\_

While streaming: Compute  $\langle a, r \rangle$ .

To estimate  $\langle \psi, a \rangle$ :

- Compute  $\langle \psi, r \rangle$
- Estimate  $\langle \psi, a \rangle^2$  to within  $\epsilon \tau \|a\|_2^2$ .

If  $\langle \psi, a \rangle^2 \geq \tau \|a\|_2^2$ , get  $\langle \psi, a \rangle^2$  to within  $(1 \pm \epsilon)$  factor.

Good streaming cost. But, apparently, high post-processing time to compute  $\langle \psi, r \rangle$ . (We'll rectify presently...)

## Reed-Muller for Randomness \_\_\_\_\_

Use 2'd-order Reed-Muller error correcting code for pseudorandomness.

String of  $r_i$ 's has dyadic structure, compatible with Walsh functions for Haar wavelet packets.

$|r|, |\psi| = N$ ; but each has  $\log^{O(1)}(N)$ -sized description.

Can compute  $\langle r, \psi \rangle$  from descriptions in time  $\log^{O(1)}(N)$ .

## Finding Large Coefficients \_\_\_\_\_

We can estimate each large wavelet packet coefficient. Do we need to try each one?

No...

## Group Testing Coefficients \_\_\_\_\_

First

- Consider signal points (delta functions).
- Seek delta function with  $2/3$  the energy.

Estimate  $\sum_{0 \leq i < N/2} a_i^2$  and  $\sum_{N/2 \leq i < N} a_i^2$ .

Exactly one *moiety* is energetic.

Similarly, estimate

$$\sum_{i \in [0, N/4) \cup [N/2, 3N/4)} a_i^2$$

and

$$\sum_{i \in [N/4, N/2) \cup [3N/4, N)} a_i^2,$$

etc.

Can identify large coeff from  $\log(N)$  tests.

## Finding Large Coefficients, generally —

To find signal points with square at least  $\tau \|a\|_2^2$ , even if  $\tau \ll 1/2$ :

Randomly assign the signal points to  $O(1/\tau^2)$  buckets. In each bucket, with high probability:

- At most one signal point of square greater than  $\tau \|a\|_2^2$ .
- Energy of non-large signal points is at most  $(\tau/2) \|a\|_2^2$ .

(Birthday paradox:  $\leq 1/\tau^2$  pairs of energetic coefficients. Each pair ends up in the same bucket with probability  $\tau^2$ . Expected number of pairs in the same bucket  $< 1$ . Repeat for better probability.)

Wavelet packet levels are similar to delta functions.

## Finding Good Representations \_\_\_\_\_

We can identify and estimate large coefficients.  
This suffices:

Suppose  $B$  terms capture energy  $\eta \|a\|_2^2$ .

Of the  $B$ , those with  $d_j^2 < \epsilon\eta/B$  have total energy  $< \epsilon\eta$ .

We find coeffs  $d_j$  with  $d_j^2 \geq \tau = \epsilon\eta/B$ ; these capture at least  $\eta - \epsilon\eta$  of the energy.

Our estimates  $\hat{d}_j = (1 \pm \epsilon)d_j$  satisfy

$$\left\| a - \sum_{k=1}^B \hat{d}_{j_k} \psi_{j_k} \right\|_2^2 \leq (1 - \eta + \epsilon\eta) \|a\|_2^2.$$

## Algorithm, Reprise ---

The following require streaming and/or post-processing actions:

- Sketch signal ( $s_a = \langle a, r \rangle$ )
- Isolate large coefficients (bucketing)
- Identify at most  $B/(\eta\epsilon)$  coefficients  $d$  with  $d^2 \geq (\epsilon\eta/B)\|a\|_2^2$  (moiety energy)
- Find sketches  $s_\psi$  of energetic wavelet vectors  $\psi$  (recursive structure of Reed-Muller)
- Estimate  $\|a\|_2$  from  $s_a$ ; get sketch  $s_a/\|a\|_2$  for  $a/\|a\|_2$
- Estimate large coeffs from  $s_a/\|a\|_2$  and  $s_\psi$ :

$$\langle a, \psi \rangle = \|a\|_2 \left( 1 - \frac{1}{2} \left\| \frac{a}{\|a\|_2} - \psi \right\|_2^2 \right).$$

- Report best  $B$  of energetic terms.

## Setup A \_\_\_\_\_

Stream through data, construct sketch.

On query  $i$ , estimate  $a_i$  from sketch alone:

Estimate  $\log(N)$  wavelet coefficients involving  $i$ , keep large ones, estimate  $a_i$  from wavelet expansion.

Which coefficients to keep? Shrinkage bounds better than our conservative bounds?

Suppose several coeffs are time-localized at  $i$ , each with large variance.

- None is valuable for overall representation.
- Collection has small variance, useful for  $a_i$ .

Better than fixed  $B$ -term representation, even computed offline?

## Exact Recovery of Superpositions \_\_\_\_\_

Suppose  $a = \sum_{k=1}^B d_{j_k} \psi_{j_k}$  exactly.

Put  $\tau = \Omega(1/B)$ , make sketch  $s_a$  of signal.

We can recover all  $d_{j_k}$ 's, even small ones, to any desired accuracy.

E.g., if  $a = \psi_j$  then  $s_a = s_{\psi_j}$ ; no matter what  $\tau$  is used for  $s_a$ .

For, initially,  $a = \sum_{k=1}^B d_{j_k} \psi_{j_k}$ , iteratively:

- Find some  $j_k$ .
- Replace  $s_a \leftarrow s_a - \hat{d}_{j_k} s_{\psi_{j_k}}$ .

Recall: In general, even if  $a$  has good rep'ns, one can only estimate  $d_j$  from sketch.

Thus: Finding best rep'n in general is *harder on the stream* than recovering superpositions.

## Other Sketchable Query Vectors \_\_\_\_\_

Given query vector  $q$ , want answer  $\langle a, q \rangle$ .

Given compact description for  $q$  and seed  $s$  for  $r$ , want sketch  $s_q = \langle q, r \rangle$ . From  $s_q$  and  $s_a = \langle a, r \rangle$ , can estimate  $\langle q, a \rangle$  to within  $\epsilon \|q\|_2 \|a\|_2$ , additively.

We can find  $s_q$  quickly if  $q$  is

- A delta function  $\delta_i$ , requesting answer  $a_i$ .
- A range, requesting answer  $\sum_{i \in [\ell, r)} a_i$ .
- A linear function  $ci + d$ , requesting answer  $\sum_i (ci + d)a_i$ .
- A function  $f$  of a few pieces, each of which is of the above form, requesting answer  $\sum_i f(i)a_i$ .

## Redundant Dictionaries \_\_\_\_\_

More than  $N$  “basis” vectors.

Two approaches:

- vectors from interlocking subspaces—e.g., wavelet packets, after Coifman and Wickerhauser.
- almost orthonormality:  $\langle v_i, v_j \rangle = \delta_{ij} \pm m$ . (Similar to Donoho’s mutual incoherence.)

$m$  is “bias.” Typically  $m = (\eta\epsilon/B)^{O(1)}$ .

Sources of error:

- $N - B$  dropped wavelet terms
- Dropped (small) terms among the  $B$
- Coeff of  $\psi$  only approximates  $\langle a, \psi \rangle$
- Can’t rank close coeffs.
- *nonzero orthonormality bias*

## Wavelet Packets ---

We find large coefficients at each resolution level.

Problem: Coefficients in nearby levels have high correlation.

Use existing algorithms (and our variants) to get good representation.

Traditional:  $B^{O(1)}N$  time.

When run on sparse tree of  $B \log(N)$  coefficients, take time only  $(B \log(N))^{O(1)}$ .

## TOPBB

---

Let  $\text{opt}(b, \ell, r, j)$  denote the energy of the best  $b$ -term approximation for coefficients  $\ell$  through  $r - 1$  at level  $j$ , and let  $C(b, \ell, r, j)$  denote the best using just terms at level  $j$ , over approximations from *library* bases.

Then

$$\begin{aligned} \text{opt}(b, \ell, r, j) &= \max \left( C(b, \ell, r, j), \right. \\ &\quad \left. \max_{0 \leq b' \leq b} \left( \text{opt}(b', \ell/2 + N/2, r/2 + N/2, j + 1) \right. \right. \\ &\quad \left. \left. + \text{opt}(b - b', \ell/2, r/2, j + 1) \right) \right). \end{aligned}$$

(This also defines library bases.)

Can compute  $\text{opt}(B, 0, N)$  in time and space  $(B \log(N))^{O(1)}$ , if all but  $B \log(N)$  coefficients are zero.

## Energy in Almost Bases \_\_\_\_\_

$D$  an almost basis of bias  $m$ .

$\psi_i \neq \psi_j \in D$  implies

$$\begin{aligned} & \|d_i\psi_i + d_j\psi_j\|_2^2 \\ &= (1 \pm m)d_i^2 + (1 \pm m)d_j^2 + 2 \langle d_i, d_j \rangle \\ &\subseteq (1 \pm m)d_i^2 + (1 \pm m)d_j^2 \pm 2m\|d_i\|_2\|d_j\|_2 \\ &\subseteq (1 \pm m)(d_i^2 + d_j^2) \end{aligned}$$

Similarly,

$$\left\| \sum d_i\psi_i \right\|_2^2 = (1 \pm Bm) \sum_i d_i^2.$$

Make  $m \ll 1/B$ . (More constraints on  $m$  to come.)

## Large Terms Suffice

---

Summing only over  $B$  terms of a rep'n with energy at least  $1 - \eta$ ,

$$\begin{aligned} \left\| \sum_{d_i^2 < \tau \|a\|_2^2} d_i \psi_i \right\|_2^2 &= (1 \pm Bm) \sum_{d_i^2 < \tau \|a\|_2^2} d_i^2 \\ &\subseteq (1 \pm Bm) B\tau \|a\|_2^2 \\ &\subseteq \epsilon\eta \|a\|_2^2, \end{aligned}$$

if  $\tau < \frac{\epsilon\eta}{2B}$  and  $m < 1/B$ . (More constraints later.)

Thus, if some  $B$ -term representation has error  $(1 - \eta)$ , there exists some  $\leq B$ -term representation, almost as good, with all  $d_i^2 \geq \tau \|a\|_2^2$ .

## Finding Good Representations \_\_\_\_\_

Find all  $j$  with  $\langle a, \psi_j \rangle^2 \geq \tau \|a\|_2^2$ . We claim:

- At most  $\approx 1/\tau$  of these
- $\sum_k \langle a, \psi_{j_k} \rangle \psi_{j_k}$  is a good rep'n of  $a$ .

## Finding Good Representations \_\_\_\_\_

Find all  $j$  with  $\langle a, \psi_j \rangle^2 \geq \tau \|a\|_2^2$ . We claim:

- At most  $\approx 1/\tau$  of these
- $\sum_k \langle a, \psi_{j_k} \rangle \psi_{j_k}$  is a good rep'n of  $a$ .

Order  $2/\tau$   $i$ 's with biggest  $\langle a, \psi_j \rangle^2$  (padding with zeros, if there are  $< 2/\tau$  such.)

Put  $\psi_{j_k}^* = \text{proj}_{(\psi_{j_1}, \dots, \psi_{j_{k-1}})^\perp}(\psi_{j_k})$ . It follows

$$\|\psi_{j_k}^* - \psi_{j_k}\|_2 \leq m(k-1)\|\psi_{j_k}\|_2 \leq \frac{2m}{\tau}\|\psi_{j_k}\|_2.$$

Thus  $\|\psi_{j_k}^*\|_2 = \left(1 \pm \frac{2m}{\tau}\right)\|\psi_{j_k}\|_2$  and  $\langle a, \psi_{j_k}^* \rangle = \langle a, \psi_{j_k} \rangle \left(1 \pm \frac{2m}{\tau}\right)$ .

Make  $m \ll (\tau/2)^2$ . It follows that

$$\sum_{j=1}^{2/\tau} \langle a, \psi_{j_k} \rangle \psi_{j_k} \approx \sum_{j=1}^{2/\tau} \langle a, \psi_{j_k}^* \rangle \psi_{j_k}^*.$$

## A Few Incoherent Bases \_\_\_\_\_

Example:

Take every  $\log(1/m)$ 'th wavelet packet row, getting  $N \log(N) / \log(1/m)$  almost basis vectors.

Handle each wavelet packet row separately.

Can afford to touch each vector when post-processing.

## Larger Dictionaries \_\_\_\_\_

Goal:

- $\tilde{\omega}(N) \gg N$  basis vectors.
- $\tilde{O}(1)$  streaming time and space
- $\tilde{O}(N)$  or even  $\tilde{o}(N)$  post-processing time.

(Hiding factors of  $\frac{\log(N)B \log(1/\delta)}{\eta\epsilon}$ .)

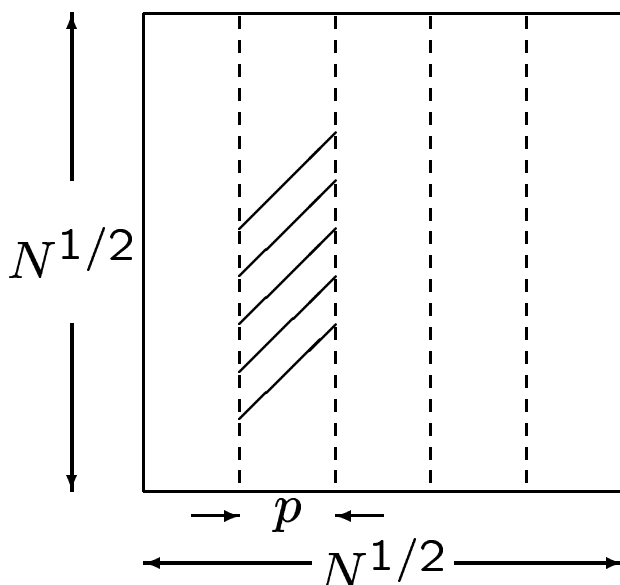
Useful even on stored data.

For  $\approx N$  time but  $\gg N$  vectors, must handle entire dictionary at once. (Can't touch each vector.)

# Segmentlets

---

After Donoho and Huo's edgelets.



- Typically  $p \approx \frac{B}{\eta\epsilon}$
- Integral slope,  $-N^{1/2}/p < \text{slope} < N^{1/2}/p$
- segmentlets “wrap around” at most once
- $N^{3/2}/p^2$  total segmentlets
- Bias  $2/p$ .
- Also incoherent with
  - horiz'l-strip segmentlets of slope  $\neq 1$
  - segmentlets with parm.  $p^j$ ,  $j = 2, 3, \dots$
  - 2-D Haarlets at scale  $p^j$ ,  $j = 0, 1, 2, 3, \dots$   
(including delta functions)

## Finding Good Segmentlet Representations

---

- $\approx N^{3/2}$  almost-basis vectors
- Streaming space and per-item time  $\tilde{O}(1)$ .
- Find energetic strips by estimating moiety energies, with cost  $\tilde{O}(1)$ .
- Search strips exhaustively; post-processing time  $\tilde{O}(N)$ .

Alternatively:

- $N^{1/4}$  slopes
- $N^{5/4}$  segmentlets
- post-processing time  $\tilde{O}(N^{3/4})$ .

Total time  $\tilde{O}(N)$ .

## Conclusions

---

When a good wavelet representation exists, we find one almost as good, on streamed data, using total space, per-item time, and post-processing time

$$\left( \frac{\log(N) B \log(1/\delta)}{\eta \epsilon} \right)^{O(1)},$$

$N$ : signal length

$B$ : number of terms

$\eta$ : energy guarantee

$\epsilon$ : energy fraction lost

$\delta$ : failure probability

Weakest possible assumption.

Useful in multiparty and periodic data settings.

We find good representations over almost bases of size  $N^{3/2}$  in total time just over  $N$ .

## Technical Mechanisms \_\_\_\_\_

- Wavelets on streams
- Randomized algorithms
- Approximation to best representation (not just approximation to signal)
- Only work on signals with good representations
- Almost bases

## Future Work

---

- General compactly-supported wavelets
- Shrinkage in Setup A
- Natural, useful, large, and low-cost redundant dictionaries
- Count and shave factors of  $\log(N)$ .

Paper Available \_\_\_\_\_

...soon. Try:

<http://www.research.att.com/~mstrauss/>

around May 1.