ORDER, INFORMATION & STREAMS

S. Guha
UPenn

Data Streams Model

- We are given a sequence of input $x_1,\ldots,x_i,\ldots,x_m$ and have to compute some function f
- Computation proceeds in passes
- Space is restricted
- Any x_i not explicitly remembered: inaccessible in the same pass

The significance of the model

- It is a model which treats “random access” as a resource.
  - The effect of the order of the input on computing a function.
  - The information we need to pass around, specially in multi-pass algorithms.

Two interpretations

- What does the stream encode
  - Whole objects: median finding
  - Updates: computing distances
- Two oldest problems in streaming…
  - (with some retro interpretation)
  - [80], [85]

Types of Order

- Adversarial
- Random
- Sorted
- Aggregated (updates)
  - Update is $\ldots\langle u, \delta^u_j\rangle_j\ldots$
  - Aggregated is $\ldots\langle u, \Sigma_j\ \delta^u_j\rangle\ldots$
- Sorted Aggregated … (time series)

- … random access $\Rightarrow$ we control the structure

Median finding

- Munro Paterson '78
  - $O\sim(n^{1/p})$ space, p passes
  - Det. Lower bounds
  - $O(\sqrt{n})$ space for 1 pass random order
  - Lower bound for "algorithms which store contiguous…"

  - Conj. $O(\log \log n)$ pass polylog space median finding algorithm exists

Approximation

- Manku, Rajagopalan, Lindsay
- Greenwald, Khanna
- …
- $O(1/\varepsilon)$ space for $\pm\ \varepsilon$ n

- ~ Munro Paterson type tradeoff
- ~ $O(1/\varepsilon^{1/p})$ space for $\pm\ \varepsilon$ n in p passes
- ~ Chang, Kannan 05…first $\Omega()$ result
  - (for a different problem)

Is there an $\Omega()$?

- Why do we care?
- Usual reasons …

- [Guha, McGregor 06] Random order
  - Polylog space $\pm$ ($\sqrt{}$n) $\log^c$ n error, 1 pass
  - O(log log n) passes suffice

- Ω $\Rightarrow$ Exponential Separation!

There is an $\Omega()$.

- Ongoing work …
- Indexing
  - Alice has $\sigma \in \{0,1\}^n$
  - Bob has j
  - Compute $\sigma[j]$
  - $\Omega(n)$ communication …

- Alice creates a stream …$2i+\sigma[i]$…
- Bob adds n-j 0's and j copies of 2(n+1)

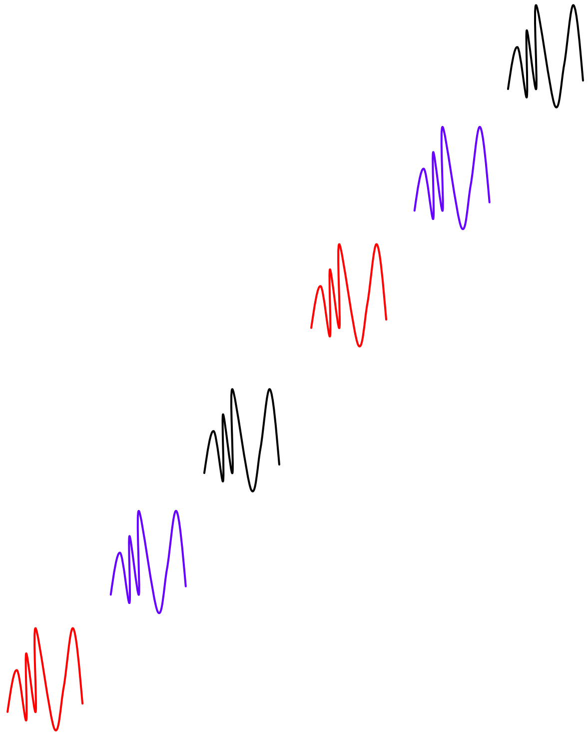
Round Elimination Lemma

- Bro-Miltersen, Nisan, Safra, Wigderson
- Communication problem $F(x,y)$
- Define P_F
  - Alice has $x_1, x_2 \ldots, x_m$
  - Bob has y, i
  - Compute $F(x_i, y)$

- Great protocol for $P_F \Rightarrow$ Good protocol for F
- If P_F is self reducible, i.e. similar to F, then..

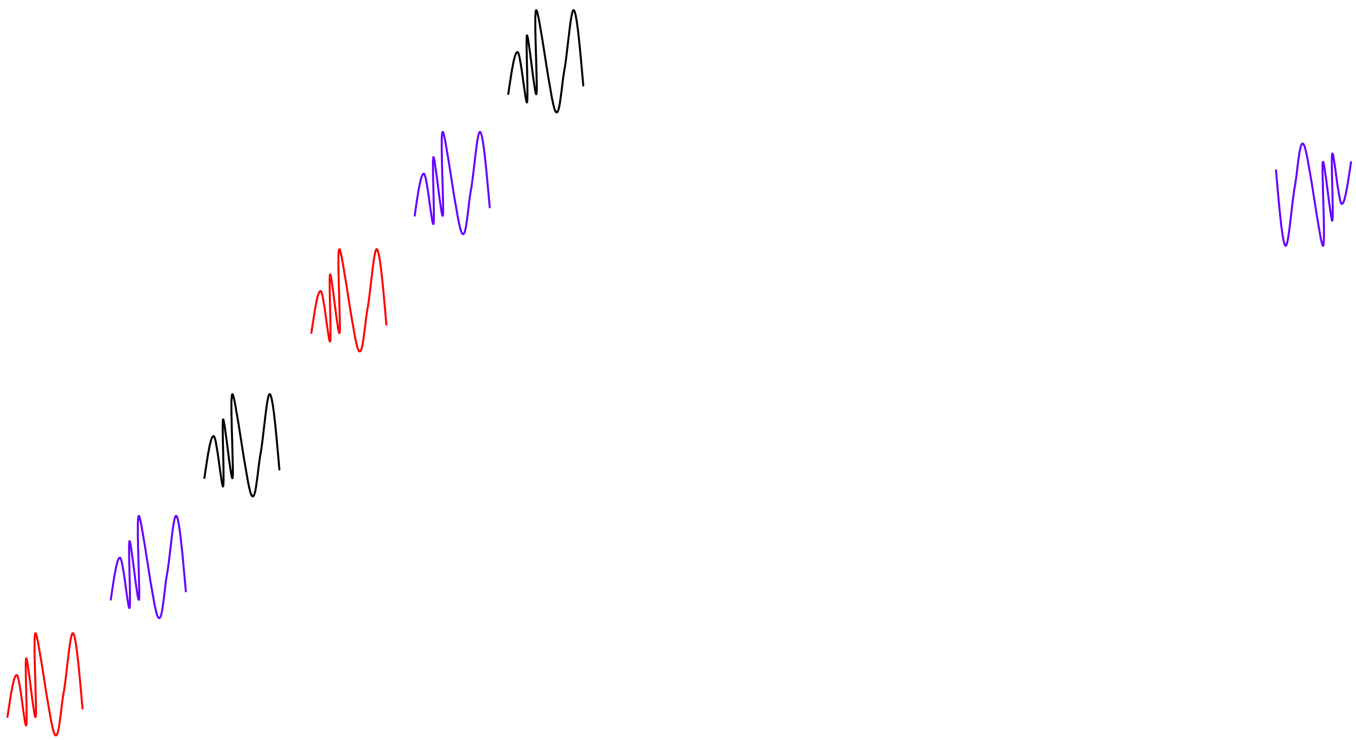
Median is self reducible

- Alice creates



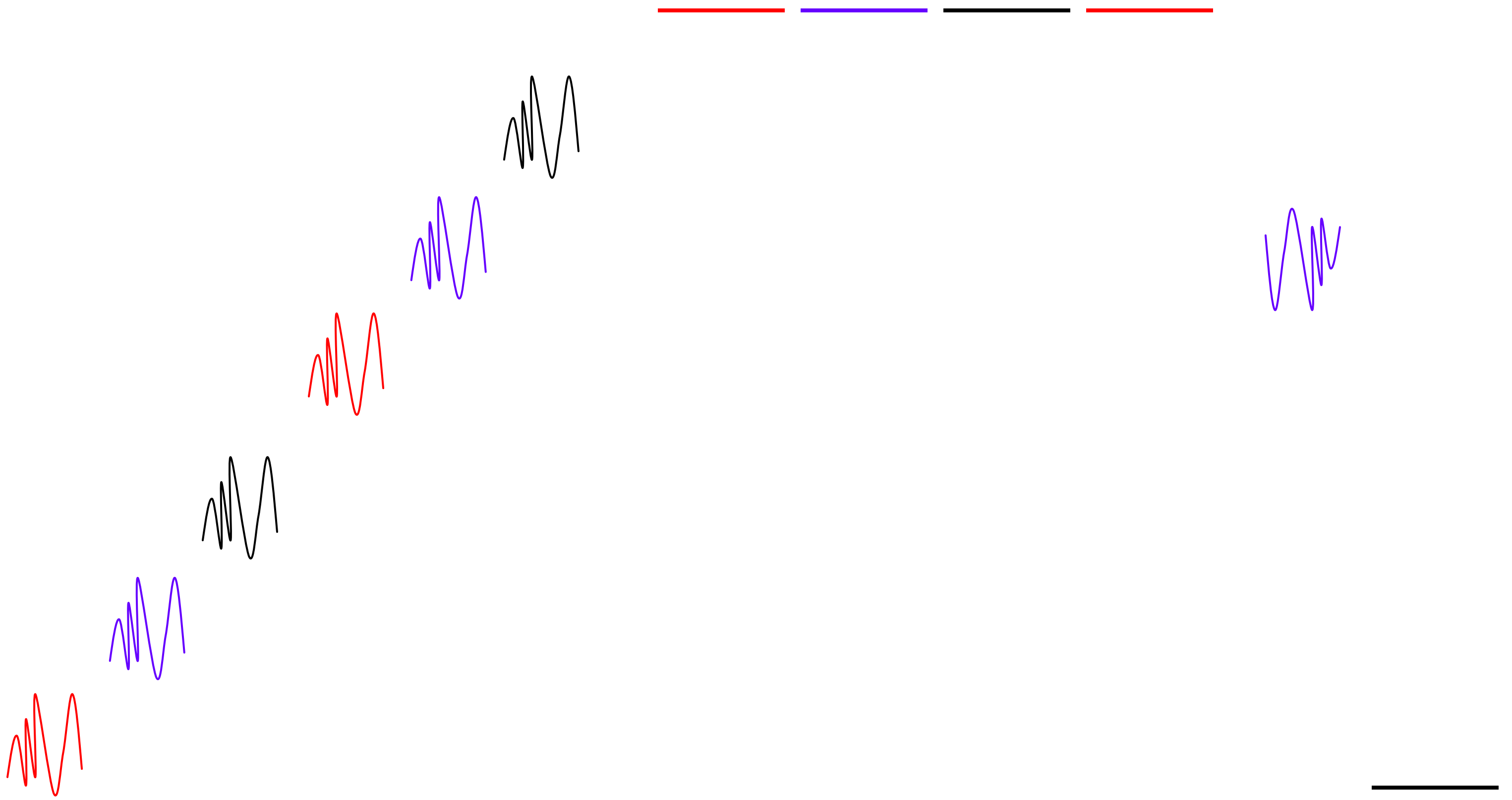
Median is self reducible

- Bob adds



Median is self reducible

- Bob adds



Result …

- $\Omega(n^{1/(2p-1)})$ space p passes
- $\Rightarrow$ Exponential Separation in random and adversarial order

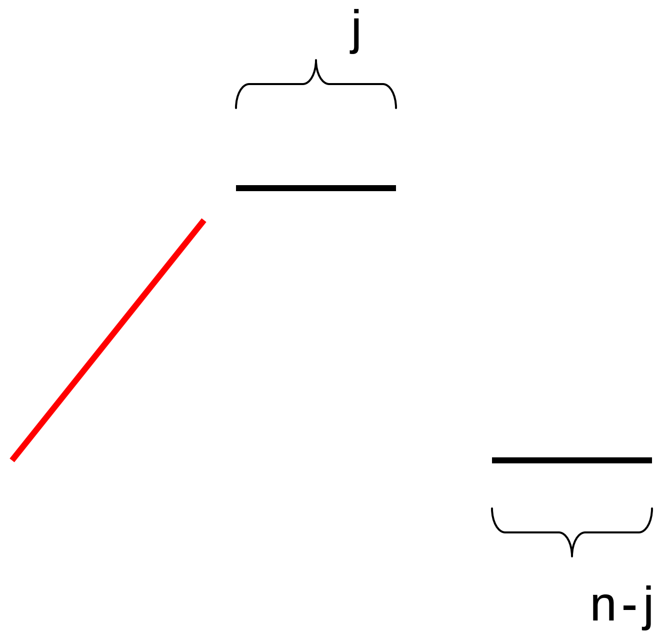
- What about other orders?
- Sorted?

Order of Medians

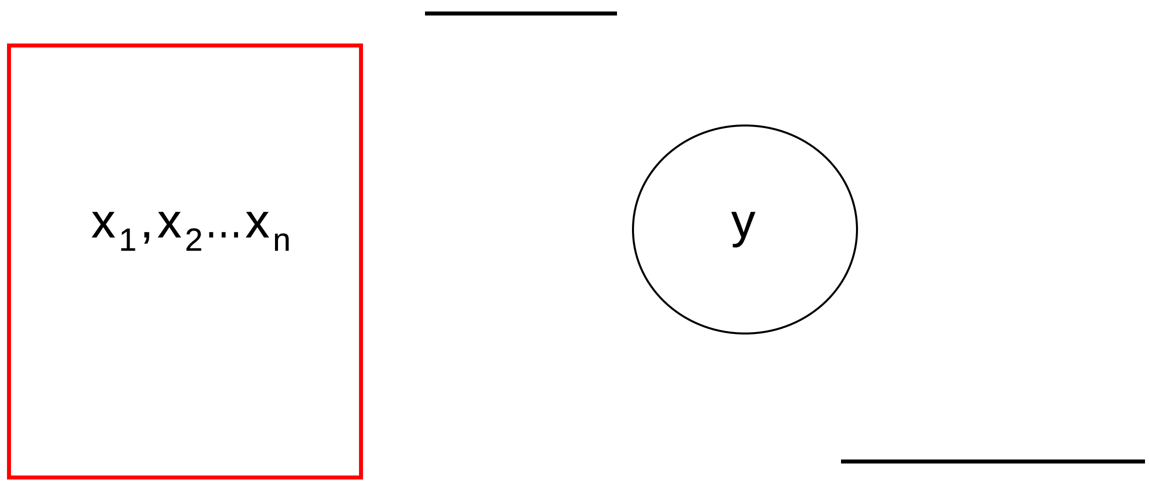
- One pass is hard – if we do not know length of stream
- Two pass is trivial.

- Variations of sorted order…
  - Bitonic?
  - Two increasing sequences?

1000 Words



1001 Words



1001 Words

What about the upper bounds?

- Median of two sorted sequences
- Emulate $O(n^{1/(2p-1)})$ protocol for Alice and Bob
  - Alice sends $O(n^{1/3})$ quantiles.
  - Bob locates the position of median
  - Sends back $O(n^{1/3})$ quantiles of that region + above, below, etc.
  - The number of candidates is now $O(n^{1/3})$
  - Alice sends $O(n^{1/3})$ numbers to Bob
  - Bob computes their rank (and of the $O(n^{1/3})$ elements he has) announces the answer

What about the upper bounds?

- Median of two sorted sequences
- The critical operation.
  - Bob locates the position of median
  - Sends back $O(n^{1/3})$ quantiles of that region + above, below, etc.

- $O(n^{1/(2p-1)})$ is tight for bitonic seq.

Adversarial Order?

- Can we do better than Munro-Paterson?
- No.

- How?
- Round Elimination does not work.

Pointer Chasing

- Alice and Bob has a function f,g resp. over [n]

- They want to compute f(g(f(g……(1))
- k alterations
- Nisan and Wigderson : $\Omega(n/k^2)$ space

- But k/2 passes … each pass has both f,g

Multiparty Pointer Chasing

- K+1 players
- Functions over [m]

- Compute $f_1(f_2(f_3(f_4......(1))$
- Consider “blowing up” the tree
- Each of $P_1, P_2, \ldots, P_k$ “anticipate” the value coming in.
- P_{k+1} dumps $f_{k+1}(1)$ to the stream
- Why medians?

Easy

- One alternation pointer chasing is Indexing.
- (slightly modified version of) Old reduction
- $\Omega(n^{1/k})$ lower bound

Interestingly…

- We have a result which separates streaming and communication complexity.

Types of Order

- Adversarial
- Random
- Sorted
- Aggregated (updates)
  - Update is $\ldots\langle u,\delta^u_j\rangle_j\ldots$
  - Aggregated is $\ldots\langle u,\Sigma_j\ \delta^u_j\rangle\ldots$
- Sorted Aggregated … (time series)

- … random access $\Rightarrow$ we control the structure

Distances between 2 streams

- Alon, Matias & Szegedy ℓ_k for $k \geq 2$ …
- Feigenbaum, Kannan, Strauss & Vishwanathan ℓ_1 but in an "aggregate model" $\Rightarrow$ … (i,# of packets) …
- Indyk ℓ_k for $1 \leq k \leq 2$ …
- Tight results for $k \geq 3$ have since been achieved…

Random Projections

- [Johnson, Lindenstrauss] 1984
- Given a matrix A whose elements are iid Gaussian, and any vector x, with high prob.

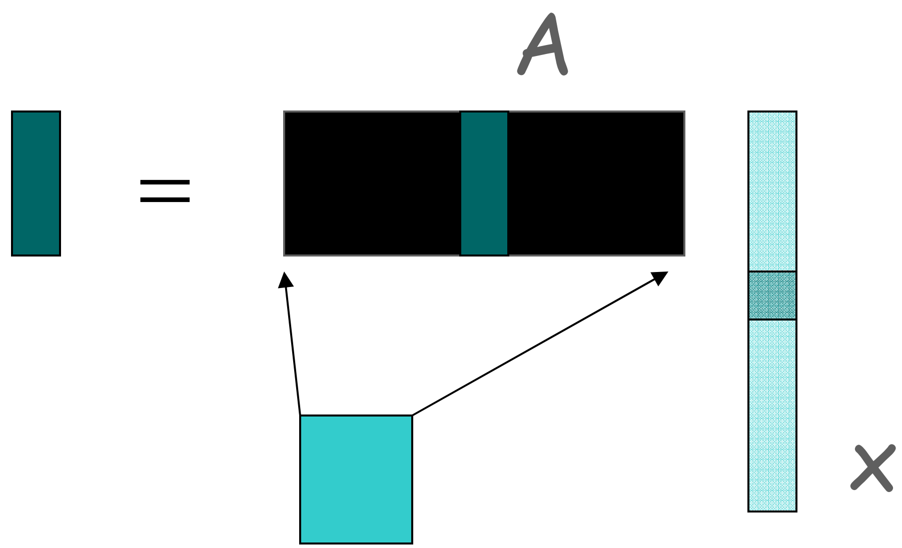
$$\|x\|_2 \le \|Ax\|_2 \le (1+\varepsilon)\|x\|_2$$

if $x \in R^n$ then $A \in R^{n \times O(\log n)}$
$\Rightarrow Ax \in R^{O(\log n)}$.

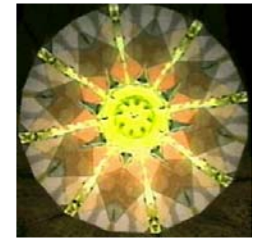
Dimensionality reduction, nearest nbr searches.

What it achieves

- Computes Norm when elements arrive out of order.



Note: A proof that such a pseudorandom generator exists is Necessary – and is not always easy.



A Kaleidoscope of questions

Which other distances are approximable?

What property of a distance makes it approximable?

You guessed it.

It's the order in which a stream arrives – and the information that comes with it.

A peek of things to come

- That's probably it folks, for update streams.
- Aggregate streams – different story.

A real Kaleidoscope of questions

- You may also ask: For what “popular” measure do we learn something new?
- Understanding is not a popularity contest.
- And popular with whom?

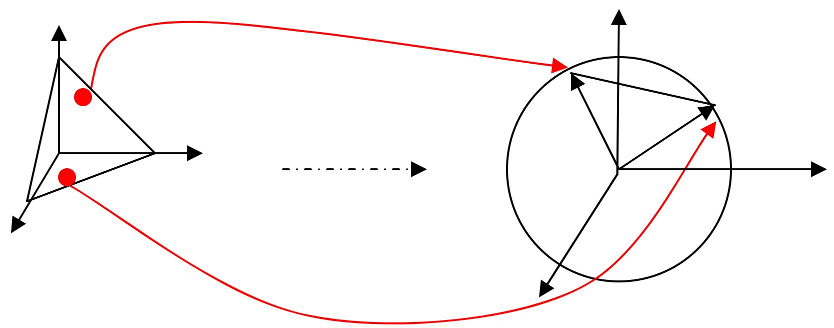
An Example

$D^2 = \Sigma_i (\sqrt{x_i} - \sqrt{y_i})^2$

(squared) Hellinger distance

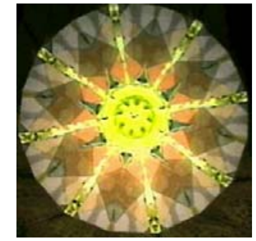
Easy in "aggregate" model

What about updates?



$\Sigma_i \sqrt{|x_i - y_i|}$ is easy (1/2 stable distribution)

A Kaleidoscope of questions



- What measures of distances are meaningful for distributions ?
  - Hypothesis testing:
    - f-divergences or Ali-Silvey-Cziszar divergences
  - Mathematical programming:
    - Bregman divergences
- Model "Risk" etc.,

Divergences

- f-divergences:
  - Pick a j from x and consider the expected likelihood $D_f(x,y)=E_{x,j}\ f(y_j/x_j)$ provided $f(1)=0$, f convex…

? ○ $KL(x,y) = \sum_j x_j \log (x_j/y_j) \Rightarrow f(u)=- \log u$

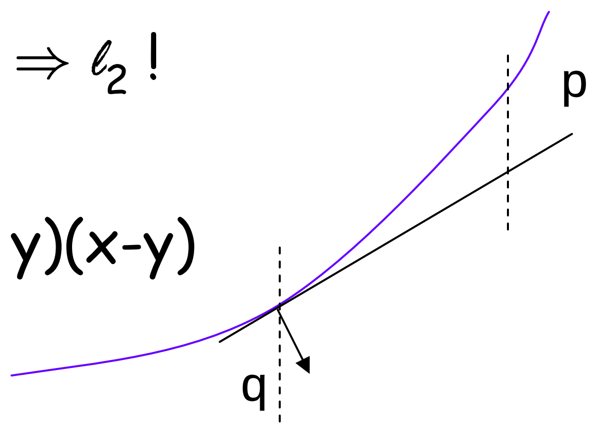
? ○ Hellinger^2 = $\sum_j (\sqrt{x_j} - \sqrt{y_j})2 = \sum_j x_j (1-\sqrt{(y_j/x_j)})^2$ or $f(u)=(1-\sqrt{u})^2$.

☺ ○ $\ell_1 = \sum_j | x_j - y_j | = \sum_j x_j | 1 - (y_j/x_j)|$ or $f(u)=|1-u|$

- Also arises from loss functions in learning …

Bregman Divergences

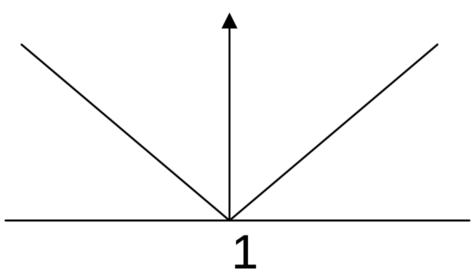
- Potential field F
- Convex F
  - $F(x)=x^2 \Rightarrow$
    - $B(x,y)=x^2-y^2-2y(x-y)=(x-y)^2 \Rightarrow \ell_2$!
  - $F(x)=x \lg x \Rightarrow$
    - $B(x,y)=x \lg x - y \lg y - (1+\lg y)(x-y)$
    $= x \lg (y/x) - x + y$
    $\Rightarrow$ Gen. KL div



$$B_F(\mathbf{p},\mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - (\nabla F(\mathbf{q})) \circ (\mathbf{p}-\mathbf{q})$$

Consequence (1)…

- If f',f" exist … f-divergences cannot be approximated in update streams
  - ℓ_1 is the ONLY f-divegence
  - We now know exactly why the other divergences do not work.



Consequence (2) …

- Bregman: If F″ vanishes or diverges polynomially at the nbd of 0 $\Rightarrow$ inapproximable.
- Note F″=constant for ℓ_2

The takeaway

- Any distance measure which is decomposable & $\phi(x_i, y_i)$ is such that it shrinks or increases even when $x_i - y_i$ is constant.
- It's the order.