Full Paper Link

Recovery from Non-Decomposable Distance Oracles

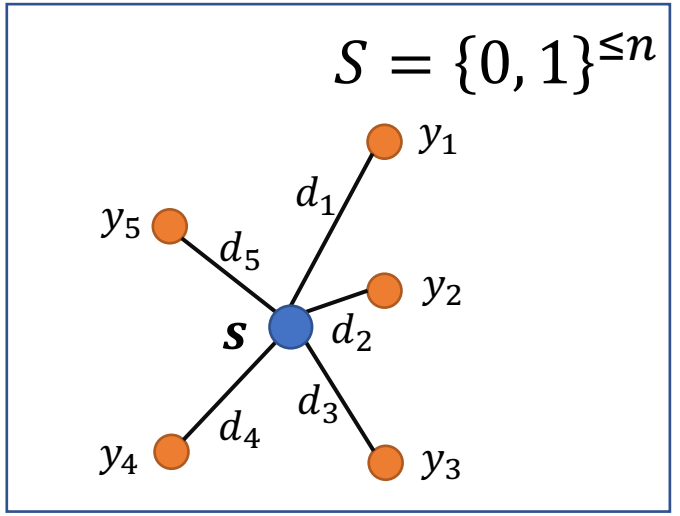
Zhuangfei Hu, Xinda Li, David P. Woodruff,
Hongyang Zhang, Shufan Zhang

University of Waterloo, Carnegie Mellon University

ITCS
Innovations in Theoretical Computer Science
ITCS 2023, MIT

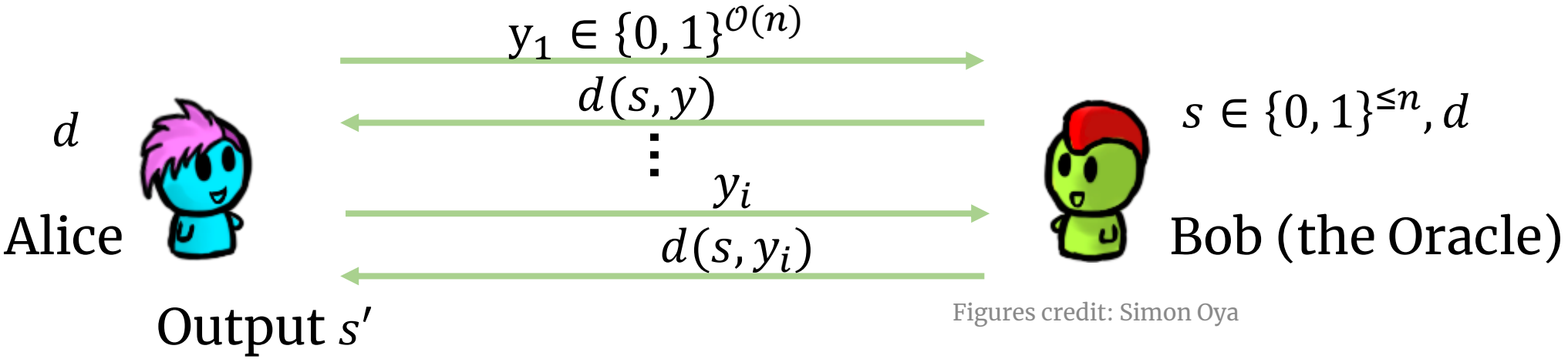
Recovery from Distance Queries

- Imagine a high-dimensional space (e.g., a hypercube), you are asked to locate an unknown point $s \in \{0,1\}^{\leq n}$
  - The knowledge you can get is a set of other points $y_i \in \mathcal{Y}$ and the distance $\mathcal{D} \coloneqq \{d_i : d(s, y_i)\}$.
  - What is the minimal size of the set $\mathcal{Y}$?



Recovery from Distance Queries

- Another way to describe this is as a 2-player game:
  - Bob plays as an oracle, embedded with sequence s and distance d.
  - Alice guesses the unknown s, by making queries to Bob.
  - Queries can be either adaptive or non-adaptive.
  - What is the query complexity for Alice to win the game?



Figures credit: Simon Oya

Prior Work

- On Hamming distance:
  - Coin-weighing [COLT'09], Group testing [COLT'20]
- On ℓ_p norm:
  - Mastermind [Approx'19]
- Other work on M-estimators [JMLR'14] [SODA'15]

- All these distances are 'decomposable':
  - $d(s, y) = \sum_i f(s_i, y_i)$ for some function f.
- The problem for 'non-decomposable' distance is not studied.

Non-Decomposable Distances

- A large class of distances:
  - Edit Distance
  - DTW
  - Fréchet Distance
  - Earth Mover Distance
  - Ulam Distance
  - Cascaded Norm (i.e., ℓ_p of ℓ_q), etc.

- This presentation will be mainly on DTW distance.

Non-Decomposable Distances

- Dynamic Time Warping (DTW)
  - Runs and Expansion
    - Runs: substring containing a single repeated character.
    - $\#runs$: the number of runs in a sequence.
    - Expansion: extending runs in a sequence.
    - E.g., $x = 010110$, $\#runs(x) = 5$, $\bar{x}_1 = 0110110, \bar{x}_2 = 0100110$.
  - Cost (ℓ_1): $||\bar{x} - \bar{y}||_1$, $\bar{x}$ and $\bar{y}$ are expansions of x and y.
    - E.g.,$||\bar{x}_1 - \bar{x}_2||_1 = 1$.
  - $d_{DTW} = \min_{(\bar{x},\bar{y})} ||\bar{x} - \bar{y}||_1$
    - E.g., $x = 010110, y = 011010$

Main Contributions

***Covered in this presentation**
******Refer to the full paper

Understanding the non-decomposable recovery problem
- Existence of indistinguishable sequences
- Three levels of recovery guarantees
- Lower bounds for recovery

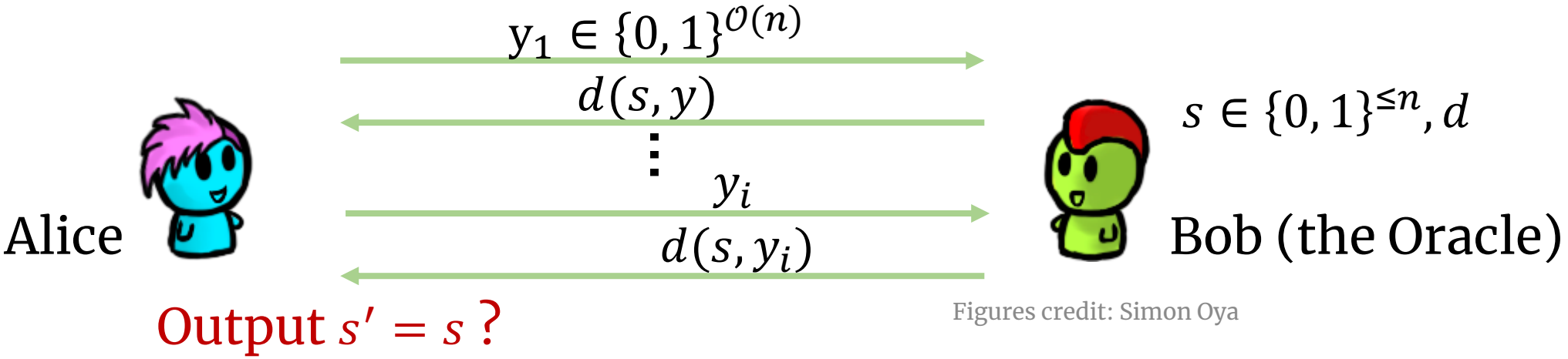
Adaptive strategies:
- General framework for all distances (sub-optimally)
- Adaptively querying edit and DTW oracle (optimally)

Non-adaptive strategies:
- Edit distance (optimally, with 1 extra char; sub-optimally, without extra char)
- DTW distance (optimally, with 2 extra chars; sub-optimally, with 1 extra char)
- Fréchet distance (optimally)

Application and open problems

Existence of Indistinguishable Sequences

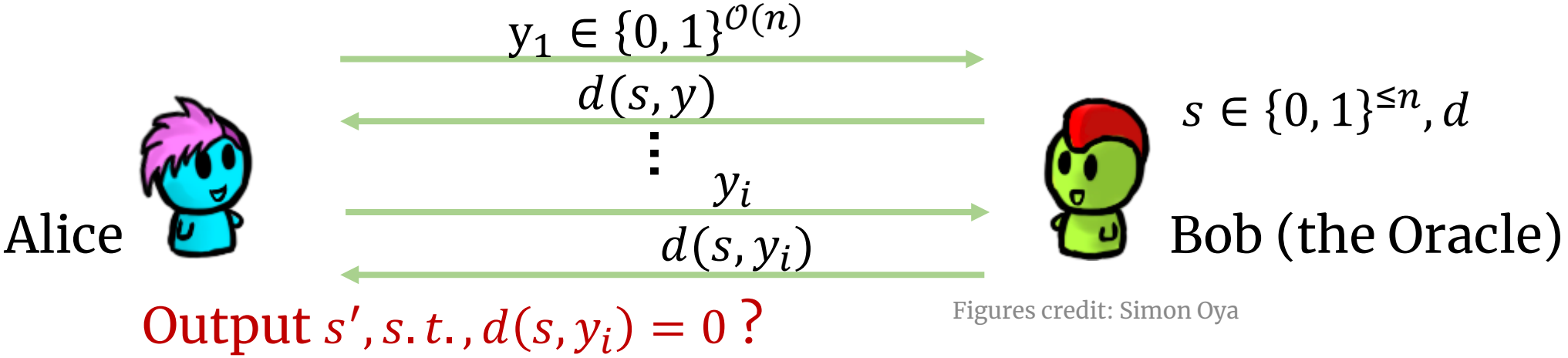
- Let's revisit the problem…
- What does it mean by "Alice wins the game"?
  - Ideally, we want $s = s'$, which means "exact sequence recovery".
  - However, this is not possible on some non-decomposable distances.



Figures credit: Simon Oya

Existence of Indistinguishable Sequences

- An example is DTW distance:
  - If $s = 010110$, when querying $y_i = 011010$, $\rightarrow d_{DTW}(s, y_i) = 0$.
  - Alice can hardly get more information about s.
  - In fact, s *and* y_i are not distinguishable by any binary seqs (Theorem 5.1 in our paper).



Figures credit: Simon Oya

Three Levels of Recovery Guarantees

- From strong to weak:
- 1. Recover the exact sequence
- 2. Recover the equivalence class
  - If x *and* y are indistinguishable to any queries, they are in the same "equivalence class".
- 3. Recover sequence with zero distance to input

Technical Results

*Covered in this presentation
**Refer to the full paper

Table 1: Summary of our results for recovering arbitrary input sequences of length n under the constraint that the query length is of $\mathcal{O}(n)$. LB: Lower Bound. #EC: Number of Extra Characters.

Oracle	Query Complexity	LB	Adaptive?	#EC	Level of Recovery	Positions
Edit	$n + \log n + c$ or $n + 2$	$\tilde{\Omega}(n)$	Adaptive	0	Exact sequence	Theorems 3.2&3.3
Edit	$n + 1$	$\tilde{\Omega}(n)$	Non-adaptive	1	Exact sequence	Theorem 4.2
Edit	n^2	$\tilde{\Omega}(n)$	Non-adaptive	0	Exact sequence	Theorem 4.4
(p-)DTW	$n + 1$	$\tilde{\Omega}(n)$	Adaptive	1	Exact sequence	Theorem 3.4
(p-)DTW	n	$\Omega(n)$	Non-adaptive	0	Equivalent class	Theorem 5.5
(p-)DTW	$n^2 + n$	$\tilde{\Omega}(n)$	Non-adaptive	1	Exact sequence	Theorem 5.7
(p-)DTW	$n + 2$	$\tilde{\Omega}(n)$	Non-adaptive	2^*	Exact sequence	Theorem 5.8
Fréchet	$2n - 1$	$2n - 1$	N/A†	0^{**}	Equivalent class	Theorem 6.3
Any distance	poly(n)	-	Adaptive	0	Zero distance to input	Theorem 3.1

† For both adaptively and non-adaptively querying the Fréchet distance oracle, the optimal bound on the query complexity is $2n - 1$.

* Increasing #EC from 2 to an arbitrary constant cannot improve the query complexity to be better than $\tilde{\mathcal{O}}(n)$.

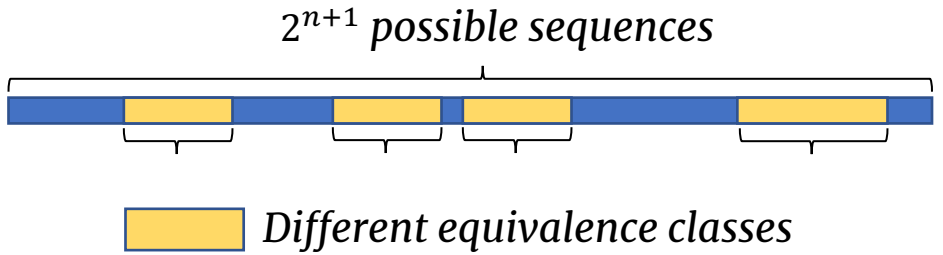
** Involving extra characters not only cannot improve the level of recovery from "equivalence class" to "exact sequence", but also cannot improve the query complexity (see Theorem 6.2).

Lower Bounds for Recovery (Exact Recover)

- $\widetilde{\Omega}(n)$ Lower bound for exact sequence recovery
  - A query result is an integer $d = \mathcal{O}(n)$.
  - The search space is $\sum_k 2^k = 2^{n+1} - 1,\ k \in [n]$.
  - By information theory, $\log_{\mathcal{O}(n)} 2^{n+1} = \Omega(\frac{n}{\log n})$ queries are needed.
- If one is allowed to use randomized algorithms…
  - By a reduction from a two-party communication game, INDEX [STOC'95], we obtain the same lower bound.

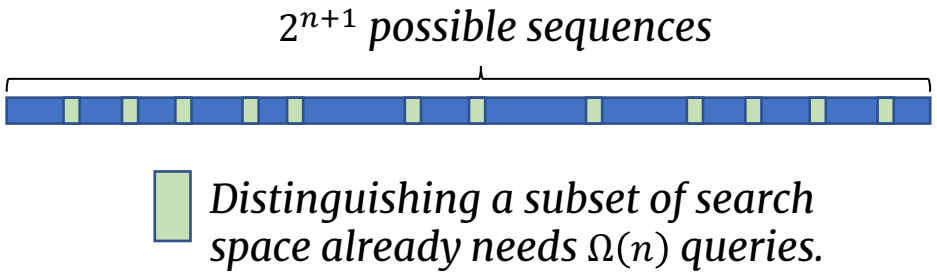
Lower Bounds for Recovery (Equivalence Class)

- $\Omega(n)$ Lower bound for equivalence class recovery
  - Due to the existence of indistinguishable sequences, you can only recover an equivalence class, if you are only allowed to make binary queries.
  - The size of the search space reduces and is hard to characterize.



Lower Bounds for Recovery (Equivalence Class)

- $\Omega(n)$ Lower bound for equivalence class recovery
  - Due to the existence of indistinguishable sequences, you can only recover an equivalence class, if you are only allowed to make binary queries.
  - The size of the search space reduces and is hard to characterize.
  - We prove this by finding a set of sequences; distinguishing them needs at least $\Omega(n)$ queries



Lower Bounds for Recovery (Proof Sketch)

- $\Omega(n)$ Lower bound for equivalence class recovery
  - DTW can be reduced to Min 1-Separated Sum (MSS) [FOCS'15]
    - $d_{DTW}(x,y) = MSS(\left(\left|x^{(2)}\right|, \dots, \left|x^{(\#runs(x)-1)}\right|\right), \frac{\#runs(x) - \#runs(y)}{2})$
    - E.g., $x = 010110, y = 010, d_{DTW}(x,y) = MSS\big((1,1,2),1\big) = 1$
    - Same MSS instances → Same DTW distance
  - Observation 1: certain pairs of $s\ and\ s'$ can only be distinguished by query with certain #runs.
    - E.g., $s = 01^3 01^3 0^3 1^3 0$ and $s' = 01^3 0^2 1^3 0^2 1^3 0^3 1^3 0 \ \rightarrow\ \#runs(q) \in [4, 10]$.
  - Observation 2: $s \in \{0,1\}^{\leq n}$ can have $\mathcal{O}(n)$ runs.
    - The $\mathcal{O}(n)$ runs can be split into $\mathcal{O}(n)$ constant intervals.
    - For each interval, there exists a pair of $s\ and\ s'$ that can only be distinguished by query with runs in this interval.

Technical Results

*Covered in this presentation
**Refer to the full paper

Table 1: Summary of our results for recovering arbitrary input sequences of length n under the constraint that the query length is of $\mathcal{O}(n)$. LB: Lower Bound. #EC: Number of Extra Characters.

Oracle	Query Complexity	LB	Adaptive?	#EC	Level of Recovery	Positions
Edit	$n + \log n + c$ or $n + 2$	$\tilde{\Omega}(n)$	Adaptive	0	Exact sequence	Theorems 3.2&3.3
Edit	$n + 1$	$\tilde{\Omega}(n)$	Non-adaptive	1	Exact sequence	Theorem 4.2
Edit	n^2	$\tilde{\Omega}(n)$	Non-adaptive	0	Exact sequence	Theorem 4.4
(p-)DTW	$n + 1$	$\tilde{\Omega}(n)$	Adaptive	1	Exact sequence	Theorem 3.4
(p-)DTW	n	$\Omega(n)$	Non-adaptive	0	Equivalent class	Theorem 5.5
(p-)DTW	$n^2 + n$	$\tilde{\Omega}(n)$	Non-adaptive	1	Exact sequence	Theorem 5.7
(p-)DTW	$n + 2$	$\tilde{\Omega}(n)$	Non-adaptive	2^*	Exact sequence	Theorem 5.8
Fréchet	$2n - 1$	$2n - 1$	N/A†	0^{**}	Equivalent class	Theorem 6.3
Any distance	poly(n)	-	Adaptive	0	Zero distance to input	Theorem 3.1

† For both adaptively and non-adaptively querying the Fréchet distance oracle, the optimal bound on the query complexity is $2n - 1$.

* Increasing #EC from 2 to an arbitrary constant cannot improve the query complexity to be better than $\tilde{\mathcal{O}}(n)$.

** Involving extra characters not only cannot improve the level of recovery from "equivalence class" to "exact sequence", but also cannot improve the query complexity (see Theorem 6.2).

General Framework for Recovery, Adaptively

- *Coordinate descent* framework for all non-decomposable distances
  - Sub-optimal, requiring $\mathcal{O}(\text{poly}(n))$ queries
  - Recovery level: Zero-distance to input

General Framework for Recovery, Adaptively

- *Coordinate descent* framework for all non-decomposable distances
  - Loss: distance function; Goal: reducing loss to zero.
  - Step operation
    - Defined as per distance: e.g, for DTW, adding/deleting/changing a character.
    - 1 step operation can reduce loss at most 1.
    - Finding a step operation to reduce loss requires at most $\mathcal{O}(\text{poly}(n))$ queries.
  - Since $\forall q, q' \in \{0,1\}, d(q,q') \leq \mathcal{O}(n)$
    - The overall query complexity is $\mathcal{O}(\text{poly}(n))$.

Adaptive Recovery Strategies

- *Coordinate Descent* is sub-optimal but fits all non-decomposable distances
- For specific instantiations, we show adaptive strategies for edit, DTW, and Fréchet with $\mathcal{O}(n), \mathcal{O}(n), \mathcal{O}(n)$ queries (cf. full paper).

Technical Results

*Covered in this presentation
**Refer to the full paper

Table 1: Summary of our results for recovering arbitrary input sequences of length n under the constraint that the query length is of $\mathcal{O}(n)$. LB: Lower Bound. #EC: Number of Extra Characters.

Oracle	Query Complexity	LB	Adaptive?	#EC	Level of Recovery	Positions
Edit	$n + \log n + c$ or $n + 2$	$\tilde{\Omega}(n)$	Adaptive	0	Exact sequence	Theorems 3.2&3.3
Edit	$n + 1$	$\tilde{\Omega}(n)$	Non-adaptive	1	Exact sequence	Theorem 4.2
Edit	n^2	$\tilde{\Omega}(n)$	Non-adaptive	0	Exact sequence	Theorem 4.4
(p-)DTW	$n + 1$	$\tilde{\Omega}(n)$	Adaptive	1	Exact sequence	Theorem 3.4
(p-)DTW	n	$\Omega(n)$	Non-adaptive	0	Equivalent class	Theorem 5.5
(p-)DTW	$n^2 + n$	$\tilde{\Omega}(n)$	Non-adaptive	1	Exact sequence	Theorem 5.7
(p-)DTW	$n + 2$	$\tilde{\Omega}(n)$	Non-adaptive	2^*	Exact sequence	Theorem 5.8
Fréchet	$2n - 1$	$2n - 1$	N/A†	0^{**}	Equivalent class	Theorem 6.3
Any distance	poly(n)	-	Adaptive	0	Zero distance to input	Theorem 3.1

† For both adaptively and non-adaptively querying the Fréchet distance oracle, the optimal bound on the query complexity is $2n - 1$.

* Increasing #EC from 2 to an arbitrary constant cannot improve the query complexity to be better than $\tilde{\mathcal{O}}(n)$.

** Involving extra characters not only cannot improve the level of recovery from "equivalence class" to "exact sequence", but also cannot improve the query complexity (see Theorem 6.2).

Equivalence Class Recovery from DTW Oracle, Non-adaptively

- Recall that, sequences $s = 010110\ and\ s' = 011010$ cannot be distinguished by any binary queries…
- We have shown an $\Omega(n)$ lower bound for equivalence class recovery.
- Now we show the construction of $\mathcal{O}(n)$ queries so that upper bound matches the lower bound (see paper for full proof).

$$z_i = \begin{cases} 0^n,\ i = 1; \\ 0^n 1 (01)^{m-1} 0^n,\ i = 2m+1; \\ 0^n (10)^{m-1} 1^n,\ i = 2m, \end{cases} \qquad o_i = \begin{cases} 1^n,\ i = 1; \\ 1^n 0 (10)^{m-1} 1^n,\ i = 2m+1; \\ 1^n (01)^{m-1} 0^n,\ i = 2m, \end{cases}$$

Equivalence Class Recovery from DTW Oracle, Non-adaptively (Proof Sketch)

- By contraposition:
  - A sequence cannot be distinguished by z_i and o_i ⇒ it is not distinguishable by any binary sequences.
- Sequences cannot be distinguished by z_i and o_i have features:
  - Same number of runs (≥ 3); same consecution of runs.
  - Same first run and last run.

$$z_i = \begin{cases} 0^n, \ i = 1; \\ 0^n 1 (01)^{m-1} 0^n, \ i = 2m+1; \\ 0^n (10)^{m-1} 1^n, \ i = 2m, \end{cases} \qquad o_i = \begin{cases} 1^n, \ i = 1; \\ 1^n 0 (10)^{m-1} 1^n, \ i = 2m+1; \\ 1^n (01)^{m-1} 0^n, \ i = 2m, \end{cases}$$

Equivalence Class Recovery from DTW Oracle, Non-adaptively (Proof Sketch)

- Using these features, we can apply the DTW-MSS reduction [FOCS'15] to analyze all possible pairs of MSS instances $MSS(s, q)$ and $MSS(s', q)$
  - In every case, we can get the same solution to MSS instances, which implies the same DTW distance.
  - For full details, please refer to our paper.

$$z_i = \begin{cases} 0^n, \ i = 1; \\ 0^n 1 (01)^{m-1} 0^n, \ i = 2m+1; \\ 0^n (10)^{m-1} 1^n, \ i = 2m, \end{cases} \qquad o_i = \begin{cases} 1^n, \ i = 1; \\ 1^n 0 (10)^{m-1} 1^n, \ i = 2m+1; \\ 1^n (01)^{m-1} 0^n, \ i = 2m, \end{cases}$$

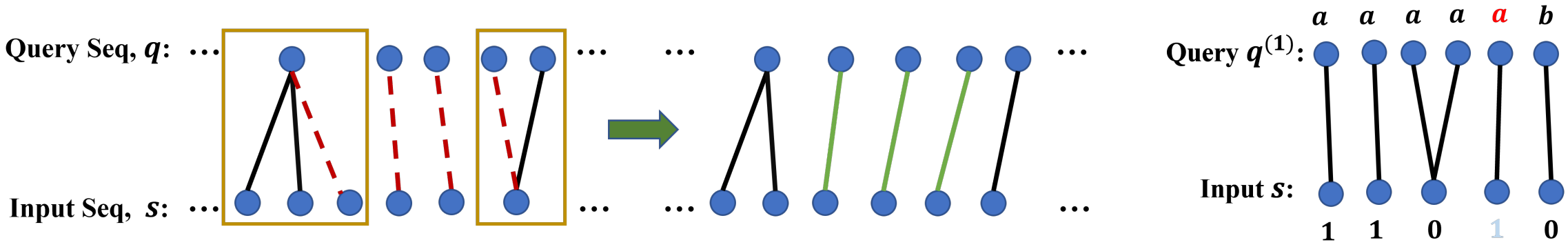
Exact Recovery from DTW Oracle, Non-adaptively

- For queries from $\{0,1\}^{\mathcal{O}(n)}$:
  - Alice can only recover to equivalence class.
  - The best query complexity is $\Theta(n)$.
- Interestingly, if Alice is allowed to query from $\left\{0,1,\frac{1}{3},\frac{2}{5}\right\}^{\mathcal{O}(n)}$, she can exactly recover the input sequence $s \in \{0,1\}^{\leq n}$...
  - Query construction: let $a = \frac{1}{3}\ and\ b = \frac{2}{5}$, define $q^{(i)} = a^{n-i}b^{i}, \forall i \in [n]$. The query set consists of all $q^{(i)}$ plus 0 and 1.
  - More generally, $a\ and\ b$ are co-prime, and $0 < b - a < a < b < \frac{1}{2}$

Exact Recovery from DTW Oracle (Proof Sketch)

$$\mathcal{Q} \coloneqq \{q^{(i)} = a^{n-i}b^{i}, \forall i \in [n]\} \cup \{0\} \cup \{1\},$$
$$a = 1/3, b = 2/5$$

- Why this set of queries works?
  - Challenge 1. How to encode information for recovery?
  - Challenge 2. How to identify the implicit alignment?
  - Let's examine some properties of these queries
- Property 1. Sequence-monotonicity → input-uniqueness
  - Let $q = q_1 q_2 \dots q_n$, we have $\min_{i \in [n]} \max\{|q_i - 0|, |q_i - 1|\} > \max_{i,j \in [n]} |q_i - q_j|$

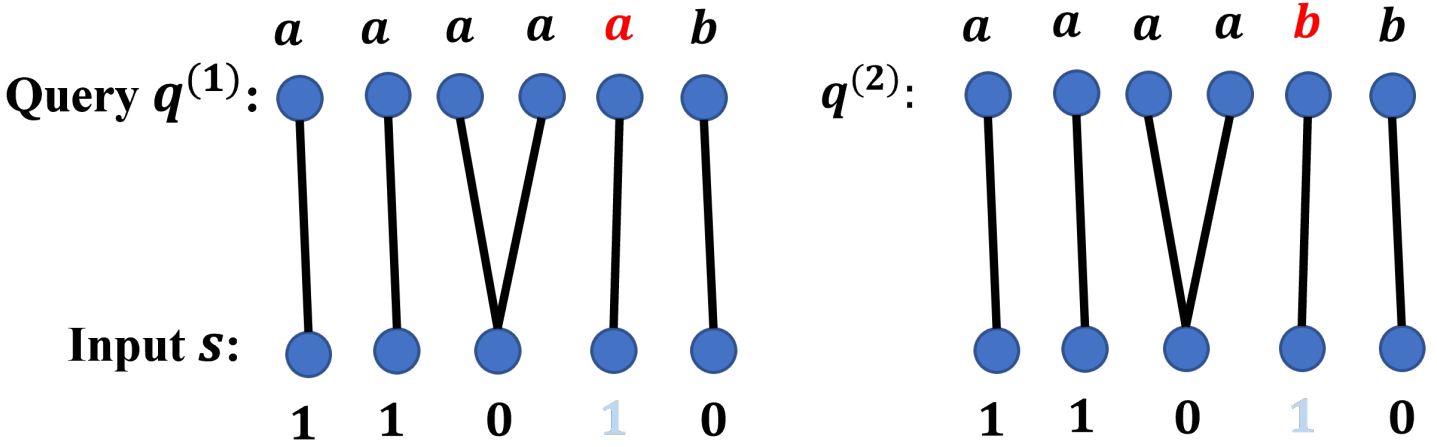


Exact Recovery from DTW Oracle (Proof Sketch)

$Q \coloneqq \{q^{(i)} = a^{n-i}b^i, \forall i \in [n]\} \cup \{0\} \cup \{1\},$

$a = 1/3, b = 2/5$

- Property 2. 0/1 preference → 0/1-uniqueness
  - All characters in $q^{(i)}$ is less than ½
  - Then every '1' in s only matches to a single character in $q^{(i)}$
- Isomorphic DTW matchings:
  - For all $q^{(i)}$, the matching between s and $q^{(i)}$ is the same.



Other Interesting Results

- Our results on DTW naturally generalize to p-DTW.
- On Fréchet Distance, there are only $\mathcal{O}(n)$ distinguishable sequences (as ℓ_∞ gives less info than ℓ_p).
- We also obtain a series of interesting results on edit distance, either optimally or near-optimally.

Application and Open Problems

- Connections to (adversarially) robust machine learning
  - Machine learning on discrete domain (e.g., NLP)
  - (Small) Perturbations are defined by edit distance, etc.
  - Certified robustness [S&P'19, ICML'19]:
    - Lipschitzness + information-preserving
  - Our non-adaptive recovery naturally yields a way to robustness.

Example 1(a).

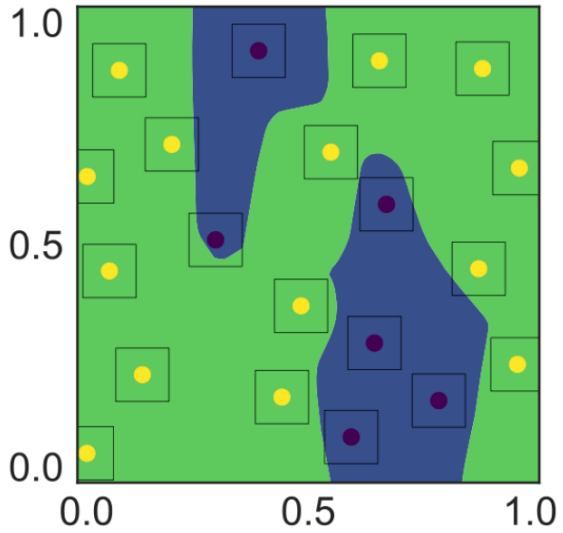
The film is weak on detail and strong on personality. → Negative (100%)

The film is wpak on detail and strong on personality. → Positive (100%)

Example 1(b).

...crypt mist in the brain. → Negative (97%)

...crypt mist in the rain. → Positive (56%)



Application and Open Problems

- Open problems
  - Closing the quadratic gap between lower and upper bounds, for edit distance, non-adaptive query complexity.
  - Solving DTW $\mathcal{O}(n)$ non-adaptive query upper bound with only 1 extra character (while we have to use 2).

- Solving any one of these two problems, we can treat him/her to an All-you-can-eat dinner, possibly Jinzakaya, in Waterloo!

Q&A

Full Paper Link

Table 1: Summary of our results for recovering arbitrary input sequences of length n under the constraint that the query length is of $\mathcal{O}(n)$. LB: Lower Bound. #EC: Number of Extra Characters.

Oracle	Query Complexity	LB	Adaptive?	#EC	Level of Recovery	Positions
Edit	$n + \log n + c$ or $n + 2$	$\tilde{\Omega}(n)$	Adaptive	0	Exact sequence	Theorems 3.2&3.3
Edit	$n + 1$	$\tilde{\Omega}(n)$	Non-adaptive	1	Exact sequence	Theorem 4.2
Edit	n^2	$\tilde{\Omega}(n)$	Non-adaptive	0	Exact sequence	Theorem 4.4
(p-)DTW	$n + 1$	$\tilde{\Omega}(n)$	Adaptive	1	Exact sequence	Theorem 3.4
(p-)DTW	n	$\Omega(n)$	Non-adaptive	0	Equivalent class	Theorem 5.5
(p-)DTW	$n^2 + n$	$\tilde{\Omega}(n)$	Non-adaptive	1	Exact sequence	Theorem 5.7
(p-)DTW	$n + 2$	$\tilde{\Omega}(n)$	Non-adaptive	2^*	Exact sequence	Theorem 5.8
Fréchet	$2n - 1$	$2n - 1$	N/A†	0^{**}	Equivalent class	Theorem 6.3
Any distance	poly(n)	-	Adaptive	0	Zero distance to input	Theorem 3.1

† For both adaptively and non-adaptively querying the Fréchet distance oracle, the optimal bound on the query complexity is $2n - 1$.

* Increasing #EC from 2 to an arbitrary constant cannot improve the query complexity to be better than $\tilde{\mathcal{O}}(n)$.

** Involving extra characters not only cannot improve the level of recovery from "equivalence class" to "exact sequence", but also cannot improve the query complexity (see Theorem 6.2).