Lower Bounds for Edit Distance and Product Metrics via Poincaré-Type Inequalities

Alexandr Andoni (Princeton University/CCI)
T.S. Jayram (IBM Almaden)
Mihai Pătrașcu (AT&T Labs)

January 17, 2010

Edit Distance

Definition

- Given two strings x and y
- $\mathrm{ED}(x, y)$ is the minimum number of insertions/deletions/substitutions to transform x into y.

Example

$$\mathrm{ED}\left(\begin{array}{l} \texttt{banana} \quad , \\ \\ \texttt{ananas} \end{array} \right) = 2$$

Edit Distance

Definition

- Given two strings x and y
- $\mathrm{ED}(x, y)$ is the minimum number of insertions/deletions/substitutions to transform x into y.

Example

$\mathrm{ED}($  ,

 $) = 2$

Edit Distance

Definition

- Given two strings x and y
- $\mathrm{ED}(x, y)$ is the minimum number of insertions/deletions/substitutions to transform x into y.

Example

ED( ,

) = 2

Applications to

- Bioinformatics,
- Text processing...

Basic Tasks

1. Compute edit distance $\mathrm{ED}(x, y)$ between strings of length d
   - Exactly: in $O(d^2/\log^2 d)$ time [Masek–Paterson'80]

Basic Tasks

1. Compute edit distance $\mathrm{ED}(x, y)$ between strings of length d
   - Exactly: in $O(d^2/\log^2 d)$ time [Masek–Paterson'80]
   - Faster? maybe up to some approximation...

Basic Tasks

1. Compute edit distance $\mathrm{ED}(x, y)$ between strings of length d
   - Exactly: in $O(d^2/\log^2 d)$ time [Masek–Paterson'80]
   - Faster? maybe up to some approximation...
   - $(\log d)^{O(1)}$ approximation in $d^{1+\varepsilon}$ time [A–Krauthgamer–Onak'??]

Basic Tasks

1. Compute edit distance $\mathrm{ED}(x, y)$ between strings of length d
   - Exactly: in $O(d^2/\log^2 d)$ time [Masek–Paterson'80]
   - Faster? maybe up to some approximation...
   - $(\log d)^{O(1)}$ approximation in $d^{1+\varepsilon}$ time [A–Krauthgamer–Onak' ??]
2. Nearest Neighbor Search: preprocess a set of n strings, so that later, given a query string, one can report its nearest neighbor
   - no exact efficient algorithms known
   - $2^{\tilde{O}(\sqrt{\log d})}$ approximation known for efficient NNS algorithms [Ostrovsky–Rabani'05]

Communication Complexity

3. The communication problem:
   - Alice has x, and Bob has y
   - they want to estimate $\mathrm{ED}(x, y)$.
   - How much information (in bits) needs to be communicated?

Communication Complexity

3. The communication problem:
   - Alice has x, and Bob has y
   - they want to estimate $\mathrm{ED}(x, y)$.
   - How much information (in bits) needs to be communicated?

Why?

- E.g., Alice and Bob have different versions of the same document

Communication Complexity

3. The communication problem:
   - Alice has x, and Bob has y
   - they want to estimate $\mathrm{ED}(x, y)$.
   - How much information (in bits) needs to be communicated?

Why?

- E.g., Alice and Bob have different versions of the same document
- Important primitive for other tasks on edit distance: implies good solutions for other problems as well.
  In fact, best bounds for (vanilla) computating edit distance and NNS problems are obtained via the communication problem!

Our Lower Bound

Alice and Bob have to compute the function

$$f_{\mathrm{ED}}(x,y) = \begin{cases} 1, & \text{if } \mathrm{ED}(x,y) \le R \quad \text{(close)} \\ 0, & \text{if } \mathrm{ED}(x,y) > \alpha R \quad \text{(far)} \end{cases}$$

for given fixed threshold R and approximation $\alpha > 1$.

Our Lower Bound

Alice and Bob have to compute the function

$$f_{\mathrm{ED}}(x,y) = \begin{cases} 1, & \text{if } \mathrm{ED}(x,y) \le R \quad \text{(close)} \\ 0, & \text{if } \mathrm{ED}(x,y) > \alpha R \quad \text{(far)} \end{cases}$$

for given fixed threshold R and approximation $\alpha > 1$.

Theorem

For some fixed threshold R, and for every approximation $\alpha < O(\log d \;/\; \log\log d)$, the communication complexity of f_{ED} is at least

$$\Omega\left(\frac{\log d \;/\; \log\log d}{\alpha}\right)$$

Our Lower Bound

Alice and Bob have to compute the function

$$f_{\mathrm{ED}}(x,y) = \begin{cases} 1, & \text{if } \mathrm{ED}(x,y) \le R \quad \text{(close)} \\ 0, & \text{if } \mathrm{ED}(x,y) > \alpha R \quad \text{(far)} \end{cases}$$

for given fixed threshold R and approximation $\alpha > 1$.

Theorem

For some fixed threshold R, and for every approximation $\alpha < O(\log d \ / \ \log\log d)$, the communication complexity of f_{ED} is at least

$$\Omega\left(\frac{\log d \ / \ \log\log d}{\alpha}\right)$$

- For approximation $\alpha = O(1)$, previously $\Omega(\log\log d)$ [A–Krauthgamer'07]
  We obtain *exponentially*–higher bound
- Lower bound works even when x, y are non-repetitive (Ulam metric)
- For Ulam metric, *close to upper bound*: $O(\log^6 d)$ communication for $\alpha = O(1)$ [A–Indyk–Krauthgamer'09]

Direct Sum Theorem

- Lower bound follows from general "direct sum" statement
- works for any metric M

Definition

Let $k \in \mathbb{N}$. The *max-product* of M is a new metric $\ell_\infty^k(M)$ where the distance between $x, y \in M^k$ is

$$\mathrm{dist}_{\infty,M}(x,y) = \max_{i=1\ldots k} \mathrm{dist}_M(x_i, y_i).$$

Direct Sum Theorem

- Lower bound follows from general "direct sum" statement
- works for any metric M

Definition

Let $k \in \mathbb{N}$. The *max-product* of M is a new metric $\ell_\infty^k(M)$ where the distance between $x, y \in M^k$ is

$$\mathrm{dist}_{\infty,M}(x,y) = \max_{i=1\ldots k} \mathrm{dist}_M(x_i, y_i).$$

Theorem (Direct Sum)

Fix approximation $\alpha > 1$, and $k \in \mathbb{N}$. If the complexity of communication problem for M is $\Omega(1)$ (an absolute constant), then the complexity of the max-product $\ell_\infty^k(M)$ is at least $\Omega(k)$.

Direct Sum in Action: Edit Distance

Theorem (Direct Sum)

Fix approximation $\alpha > 1$, and $k \in \mathbb{N}$. If the complexity of communication problem for M is $\Omega(1)$ (an absolute constant), then the complexity of the max-product $\ell_\infty^k(M)$ is at least $\Omega(k)$.

Direct Sum in Action: Edit Distance

Theorem (Direct Sum)

Fix approximation $\alpha > 1$, and $k \in \mathbb{N}$. If the complexity of communication problem for M is $\Omega(1)$ (an absolute constant), then the complexity of the max-product $\ell_\infty^k(M)$ is at least $\Omega(k)$.

- For edit distance, jump–start with:

Theorem ([A–Krauthgamer'07])

For edit distance (when $M = \mathrm{ED}$), the complexity of the communication problem is $\Omega(1)$ for approximation $\alpha_1 = \Theta(\frac{\log d}{\log\log d})$.

Direct Sum in Action: Edit Distance

Theorem (Direct Sum)

Fix approximation $\alpha > 1$, and $k \in \mathbb{N}$. If the complexity of communication problem for M is $\Omega(1)$ (an absolute constant), then the complexity of the max-product $\ell_\infty^k(M)$ is at least $\Omega(k)$.

- For edit distance, jump–start with:

Theorem ([A–Krauthgamer'07])

For edit distance (when $M = \mathrm{ED}$), the complexity of the communication problem is $\Omega(1)$ for approximation $\alpha_1 = \Theta(\frac{\log d}{\log\log d})$.

- Then, apply the direct sum theorem for $k \approx \alpha_1/100$ to obtain the lower bound for $\ell_\infty^k(\mathrm{ED})$ for approximation α_1.

Direct Sum in Action: Edit Distance

Theorem (Direct Sum)

Fix approximation $\alpha > 1$, and $k \in \mathbb{N}$. If the complexity of communication problem for M is $\Omega(1)$ (an absolute constant), then the complexity of the max-product $\ell_\infty^k(M)$ is at least $\Omega(k)$.

- For edit distance, jump–start with:

Theorem ([A–Krauthgamer'07])

For edit distance (when $M = \mathrm{ED}$), the complexity of the communication problem is $\Omega(1)$ for approximation $\alpha_1 = \Theta(\frac{\log d}{\log\log d})$.

- Then, apply the direct sum theorem for $k \approx \alpha_1/100$ to obtain the lower bound for $\ell_\infty^k(\mathrm{ED})$ for approximation α_1.
- $\ell_\infty^k(\mathrm{ED}) \approx \mathrm{ED}$ up to approximation k. Namely, we can map tuples of strings in $\ell_\infty^k(\mathrm{ED})$ to strings under edit distance (e.g., by concatenating the k strings, with some padding in between).

Proof of Direct Sum Theorem

Three steps:

Proof of Direct Sum Theorem

Three steps:

1. Show that $\Omega(1)$ communication lower bound for f_M is equivalent to a certain Poincaré-type inequality on the metric M

Proof of Direct Sum Theorem

Three steps:

1. Show that $\Omega(1)$ communication lower bound for f_M is equivalent to a certain Poincaré-type inequality on the metric M
2. The Poincaré-type inequality implies a lower bound on information complexity of a communication protocol

Proof of Direct Sum Theorem

Three steps:

1. Show that $\Omega(1)$ communication lower bound for f_M is equivalent to a certain Poincaré-type inequality on the metric M
2. The Poincaré-type inequality implies a lower bound on information complexity of a communication protocol
3. Use a direct sum theorem for the AND function of [Chakrabarti–Shi–Wirth–Yao'01, Bar-Yossef–Jayram–Kumar–Sivakumar'03]

1. Poincaré Inequalities and Protocols of Constant Communication

Fix approximation α, threshold R.

Definition (Poincaré Inequality)

- distribution η_1 on close instances ($\text{dist}_M(x,y) \le R$)
- distribution η_0 on far instances ($\text{dist}_M(x,y) > \alpha R$)
- parameters $\lambda > 0, \beta \ge 0$

A Poincaré inequality holds for M if for all $\rho : M \to \ell_2$:

$$\mathbb{E}_{(x,y)\sim\eta_1} \|\rho(x) - \rho(y)\|^2 \ge \lambda \cdot \mathbb{E}_{(x,y)\sim\eta_0} \|\rho(x) - \rho(y)\|^2 - \beta$$

1. Poincaré Inequalities and Protocols of Constant Communication

Fix approximation α, threshold R.

Definition (Poincaré Inequality)

- distribution η_1 on close instances ($\text{dist}_M(x,y) \le R$)
- distribution η_0 on far instances ($\text{dist}_M(x,y) > \alpha R$)
- parameters $\lambda > 0, \beta \ge 0$

A Poincaré inequality holds for M if for all $\rho : M \to \ell_2$:

$$\mathbb{E}_{(x,y)\sim\eta_1} \|\rho(x) - \rho(y)\|^2 \ge \lambda \cdot \mathbb{E}_{(x,y)\sim\eta_0} \|\rho(x) - \rho(y)\|^2 - \beta$$

Lemma

Suppose the communication problem for M has $\omega(1)$ communication. Then a Poincaré inequality holds for $\lambda = 1$ and $\beta = o(1)$.

Note: the converse is also true [A–Krauthgamer'07]

2. Information Complexity

- Information complexity of a function f is the minimal mutual information between the inputs (x, y) and the protocol Π (over the choice of protocols Π that correctly compute f):

$$IC(f) = \min_{\Pi} I(x, y; \Pi)$$

  Convenient to consider $I(x, y; \Pi \mid D)$ for some event D such that x and y are independently distributed when conditioned on D.

Remember, we use the partial function

$$f = f_M(x, y) = \begin{cases} 1, & \text{if } \mathrm{dist}_M(x, y) \leq R \\ 0, & \text{if } \mathrm{dist}_M(x, y) > \alpha R \end{cases}$$

2. Information Complexity

- Information complexity of a function f is the minimal mutual information between the inputs (x, y) and the protocol Π (over the choice of protocols Π that correctly compute f):

$$IC(f) = \min_{\Pi} I(x, y; \Pi)$$

Convenient to consider $I(x, y; \Pi \mid D)$ for some event D such that x and y are independently distributed when conditioned on D.

Remember, we use the partial function

$$f = f_M(x, y) = \begin{cases} 1, & \text{if } \mathrm{dist}_M(x, y) \leq R \\ 0, & \text{if } \mathrm{dist}_M(x, y) > \alpha R \end{cases}$$

Lemma

Fix some metric M, approximation α, and threshold R. If M satisfies a Poincaré inequality with $\lambda = \Omega(1)$ and $\beta = o(1)$, then $IC(f_M) \geq \Omega(1)$.

3. Direct Sum for AND Function

- Let g be the function of communication problem for $\ell_\infty^k(M)$
- Note that g is defined on M^k and

$$g(x,y) = \begin{cases} 1, & \text{if } \max_{i=1\dots k} \mathsf{dist}_M(x_i,y_i) \le R \\ 0, & \text{if } \max_{i=1\dots k} \mathsf{dist}_M(x_i,y_i) > \alpha R \end{cases} = \bigwedge_{i=1\dots k} f_M(x_i,y_i)$$

3. Direct Sum for AND Function

- Let g be the function of communication problem for $\ell_\infty^k(M)$
- Note that g is defined on M^k and

$$g(x,y) = \begin{cases} 1, & \text{if } \max_{i=1\dots k} \mathrm{dist}_M(x_i,y_i) \le R \\ 0, & \text{if } \max_{i=1\dots k} \mathrm{dist}_M(x_i,y_i) > \alpha R \end{cases} = \bigwedge_{i=1\dots k} f_M(x_i,y_i)$$

- For the AND function, we can use the following direct-sum theorem for communication complexity:

Theorem ([Chakrabarti–Shi–Wirth–Yao'01, Bar-Yossef–Jayram–Kumar–Sivakumar'03])

Let f, g be any partial functions such that $g(x,y) = \bigwedge_{i=1\dots k} f(x_i,y_i)$. Then the communication complexity of g is at least $k \cdot IC(f)$.

Direct Sum Theorem: Proof Wrap-up

Communication problem for M requires $\omega(1)$ bits of communication

Direct Sum Theorem: Proof Wrap-up

Communication problem for M requires $\omega(1)$ bits of communication

$\Longrightarrow$ Poincaré inequality for M: there exist η_0, η_1, s.t. for any $\rho : M \to \ell_2$

$$\mathbb{E}_{(x,y)\sim\eta_1} \|\rho(x) - \rho(y)\|^2 \geq \mathbb{E}_{(x,y)\sim\eta_0} \|\rho(x) - \rho(y)\|^2 - o(1)$$

Direct Sum Theorem: Proof Wrap-up

Communication problem for M requires $\omega(1)$ bits of communication

$\implies$ Poincaré inequality for M: there exist η_0, η_1, s.t. for any $\rho : M \to \ell_2$

$$\mathbb{E}_{(x,y)\sim\eta_1} \|\rho(x) - \rho(y)\|^2 \geq \mathbb{E}_{(x,y)\sim\eta_0} \|\rho(x) - \rho(y)\|^2 - o(1)$$

$\implies$ Information complexity lower bound for f_M of $\Omega(1)$:

$$IC(f_M) \geq \Omega(1)$$

Direct Sum Theorem: Proof Wrap-up

Communication problem for M requires $\omega(1)$ bits of communication

$\Longrightarrow$ Poincaré inequality for M: there exist η_0, η_1, s.t. for any $\rho : M \to \ell_2$

$$\mathbb{E}_{(x,y)\sim\eta_1} \|\rho(x) - \rho(y)\|^2 \geq \mathbb{E}_{(x,y)\sim\eta_0} \|\rho(x) - \rho(y)\|^2 - o(1)$$

$\Longrightarrow$ Information complexity lower bound for f_M of $\Omega(1)$:

$$IC(f_M) \geq \Omega(1)$$

$\Longrightarrow$ Communication complexity lower bound $\bigwedge_{i=1}^{k} f_M$ is $\Omega(k)$; equivalent to communication problem on $\ell_\infty^k(M)$.

Conclusion

- Lower bound of $\Omega(\frac{\log d \ / \ \log\log d}{\alpha})$ for communication complexity of estimating edit distance up to approximation α
- Close to the upper bound of $O(\log^6 d)$ in the case of Ulam distance (non-repetitive strings)
- Lower bound based on a geometric feature of the metric (a Poincaré-type inequality), usually used for proving non-embeddability into normed spaces

Conclusion

- Lower bound of $\Omega(\frac{\log d \ / \ \log\log d}{\alpha})$ for communication complexity of estimating edit distance up to approximation α
- Close to the upper bound of $O(\log^6 d)$ in the case of Ulam distance (non-repetitive strings)
- Lower bound based on a geometric feature of the metric (a Poincaré-type inequality), usually used for proving non-embeddability into normed spaces

Open question: upper bounds on communication complexity?

Conclusion

- Lower bound of $\Omega(\frac{\log d \ / \ \log\log d}{\alpha})$ for communication complexity of estimating edit distance up to approximation α
- Close to the upper bound of $O(\log^6 d)$ in the case of Ulam distance (non-repetitive strings)
- Lower bound based on a geometric feature of the metric (a Poincaré-type inequality), usually used for proving non-embeddability into normed spaces

Open question: upper bounds on communication complexity?

- $2^{\tilde{O}(\sqrt{\log d})}$ approximation with $O(1)$ bits [Ostrovsky–Rabani'05]

Conclusion

- Lower bound of $\Omega(\frac{\log d \ / \ \log\log d}{\alpha})$ for communication complexity of estimating edit distance up to approximation α
- Close to the upper bound of $O(\log^6 d)$ in the case of Ulam distance (non-repetitive strings)
- Lower bound based on a geometric feature of the metric (a Poincaré-type inequality), usually used for proving non-embeddability into normed spaces

Open question: upper bounds on communication complexity?

- $2^{\tilde{O}(\sqrt{\log d})}$ approximation with $O(1)$ bits [Ostrovsky–Rabani'05]
- Partial progress: $(\log d)^{O(1)}$ approximation with d^{ε} bits when $R = \tilde{\Theta}(d)$ [A–Krauthgamer–Onak]