Polylogarithmic Approximation for Edit Distance and the Asymmetric Query Complexity

Alexandr Andoni, Robert Krauthgamer, Krysztof Onak

Theorem 1.1 (Main): *For every fixed $\varepsilon > 0$, there is an algorithm that approximates the edit distance between two input strings $x, y \in \Sigma^n$ within a factor of $(\log n)^{\mathcal{O}(1/\varepsilon)}$, and runs in $n^{1+\varepsilon}$ time.*

Definition: Consider two strings $x, y \in \Sigma^n$ for some alphabet Σ, and let $\mathrm{ed}(x, y)$ denote the edit distance between these two strings. The computational problem is the promise problem known as the *Distance Threshold Estimation Problem (DTEP)*: distinguish whether $\mathrm{ed}(x, y) > R$ or $\mathrm{ed}(x, y) \leq R/\alpha$, where $R > 0$ is a parameter (known to the algorithm) and $\alpha \geq 1$ is the approximation factor. We use DTEP_β to denote the case of $R = n/\beta$, where $\beta \geq 1$ may be a function of n.

Definition: In the *asymmetric query model*, the algorithm knows in advance (has unrestricted access to) one of the strings, say y, and has only *query access* to the other string, x. The *asymmetric query complexity* of an algorithm is the number of coordinates in x that the algorithm has to probe in order to solve DTEP with success probability at least 2/3.

Theorem 1.2 (Query complexity upper bound): *For every $\beta = \beta(n) \geq 2$ and fixed $0 < \varepsilon < 1$ there is an algorithm that solves DTEP_β with approximation $\alpha = (\log n)^{\mathcal{O}(1/\varepsilon)}$, and makes βn^ε asymmetric queries. This algorithm runs in time $\mathcal{O}(n^{1+\varepsilon})$.*

For every $\beta = \mathcal{O}(1)$ and fixed integer $t \geq 2$ there is an algorithm for DTEP_β achieving approximation $\alpha = \mathcal{O}(n^{1/t})$, with $\mathcal{O}(\log^{t-1} n)$ queries into x.

Theorem 1.3 (Query complexity lower bound): *For a sufficiently large constant $\beta > 1$, every algorithm that solves DTEP_β with approximation $\alpha = \alpha(n) > 2$ has asymmetric query complexity $2^{\Omega\left(\frac{\log n}{\log \alpha + \log\log n}\right)}$. Moreover, for every fixed non-integer $t > 1$, every algorithm that solves DTEP_β with approximation $\alpha = n^{1/t}$ has asymmetric query complexity $\Omega(\log^{\lfloor t \rfloor} n)$.*

Theorem 3.1: *Let $n \geq 2$, $\beta = \beta(n) \geq 2$, and integer $b = b(n) \geq 2$ be such that $(\log_b n) \in \mathbb{N}$.*

There is an algorithm solving DTEP_β with approximation $\alpha = \mathcal{O}(b \log_b n)$ and $\beta \cdot (\log n)^{\mathcal{O}(\log_b n)}$ queries into x. The algorithm runs in $n \cdot (\log n)^{\mathcal{O}(\log_b n)}$ time.

For every constant $\beta = \mathcal{O}(1)$ and integer $t \geq 2$, there is an algorithm for solving DTEP_β with $\mathcal{O}(n^{1/t})$ approximation and $\mathcal{O}(\log^{t-1} n)$ queries. The algorithm runs in $\tilde{\mathcal{O}}(n)$ time.

Characterization of edit distance using $\mathcal{E}$-distance

For a string x, $x[s : t]$ denotes the substring of x comprising of $x[s], \ldots, x[t-1]$. The characterization may be viewed as a tree of arity b, where nodes correspond to substring $x[s : s + l]$. The root is the entire string $x[1 : n + 1]$. Let $h \overset{\mathrm{def}}{=} \log_b n \in \mathbb{N}$. Then nodes on level i for $0 \leq i \leq h$ correspond to substrings $x[s : s + l_i]$ of length $l_i \overset{\mathrm{def}}{=} n/b^i$.

Definition 3.2 ($\mathcal{E}$-distance): Consider two strings x, y of length $n \geq 2$. Fix $i \in \{0, 1, \ldots, h\}$, $s \in B_i = \{1, 1 + l_i, \ldots\}$, and a position $u \in \mathbb{Z}$.

If $i = h$, the *$\mathcal{E}$-distance* of $x[s : s+1]$ to the position u is 1 if $u \notin [n]$ or $x[s] \neq y[u]$, and 0 otherwise.

For $i \in \{0, 1, \ldots, h - 1\}$, we recursively define the *$\mathcal{E}$-distance* $\mathcal{E}_{x,y}(i, s, u)$ of $x[s : s + l_i]$ to the position u as follows. Partition $x[s : s + l_i]$ into b blocks of length $l_{i+1} = l_i/b$, starting at positions $s + jl_{i+j}$, where $j \in \{0, 1, \ldots, b - 1\}$. Then

$$\mathcal{E}_{x,y}(i, s, u) \overset{\mathrm{def}}{=} \sum_{j=0}^{b-1} \min_{r_j \in \mathbb{Z}} \mathcal{E}_{x,y}(i + 1, s + jl_{i+1}, u + jl_{i+1} + r_j) + |r_j|.$$

The $\mathcal{E}$-distance from x to y is $\mathcal{E}_{x,y}(0, 1, 1)$.

Theorem 3.3 (Characterization): *For every $b \geq 2$ and two strings $x, y \in \Sigma^n$, the $\mathcal{E}$-distance between x and y is a $6 \cdot \frac{b}{\log b} \cdot \log n$ approximation to the edit distance between x and y.*

Definition (Alternative): Consider all the matching positions of $\mathcal{E}$ during the computation. Denote by Z a vector of integers $z_{i,s}$ indexed by $i \in \{0, 1, \ldots, h\}$ and $s \in B_i = \{1, 1 + l_i, \ldots\}$, where $z_{0,1} = 1$ by convention. The coordinate $z_{i,s}$ should be understood as the position to which we match the substring $x[s : s + l_i]$. Then we define the cost of Z as

$$\operatorname{cost}(Z) \overset{\text{def}}{=} \sum_{i=0}^{h-1} \sum_{s \in B_i} \sum_{j=0}^{b-1} |z_{i,s} + jl_{i+1} - z_{i+1,s+jl_{i+1}}|.$$

Claim 3.4 (Alternative definition of $\mathcal{E}$-distance): The $\mathcal{E}$-distance between x and y is the minimum of

$$\operatorname{cost}(Z) + \sum_{s \in [n]} \mathrm{H}(x[s], y[z_{h,s}])$$

over all choices of the vector $Z = (z_{i,s})_{i \in \{0,1,\ldots,h\}, s \in B_i}$ with $z_{0,1} = 1$, where $\mathrm{H}(\cdot, \cdot)$ is the Hamming distance.

Lemma 3.5: The $\mathcal{E}$-distance between x and y is at most $3hb \cdot \operatorname{ed}(x, y)$.

Lemma 3.6: The edit distance $\operatorname{ed}(x, y)$ is at most twice the $\mathcal{E}$-distance between x and y.

Sampling Algorithm

Chernoff bound: Let $Z_i \in [0, 1]$ be n independent random variables from possibly different distributions. Let $Z = \sum_i Z_i$ and $\mu = \mathbb{E}[Z]$. Then for any $\varepsilon > 0$:

$$\mathbb{P}[Z \geq (1+\varepsilon)\mu] \leq e^{-\frac{\varepsilon^2 \mu}{2+\varepsilon}} \qquad \text{and} \qquad \mathbb{P}[Z \leq (1-\varepsilon)\mu] \leq e^{-\frac{\varepsilon^2 \mu}{2}}.$$

Hoeffding bound: Let $Z_i \in [0, 1]$ be n independent random variables from possibly different distributions. Let $Z = \sum_i Z_i$ and $\mu = \mathbb{E}[Z]$. Then for any $t > 0$, we have that

$$\mathbb{P}[Z \geq t] \leq e^{-(t-2\mu)}.$$

Definition 3.8: Fix $\rho > 0$ and some $f \in [1, 2]$. For a quantity $\tau \geq 0$, we call its (ρ, f)-*approximator* any quantity $\hat{\tau}$ such that $\tau/f - \rho \leq \hat{\tau} \leq f\tau + \rho$.

If $\hat{\tau}_1, \hat{\tau}_2$ are (ρ, f)-approximators to τ_1, τ_2 respectively, $\hat{\tau}_1 + \hat{\tau}_2$ is a $(2\rho, f)$-approximator to $\tau_1 + \tau_2$.

If $\hat{\tau}'$ is a (ρ', f')-approximator to $\hat{\tau}$, which itself is a (ρ, f)-approximator to τ, then $\hat{\tau}'$ is a $(\rho' + f'\rho, ff')$-approximator to τ.

Lemma 3.9 (Sum of random variables): Fix $n \in \mathbb{N}$, $\rho > 0$ and error probability δ. Let $Z_i \in [0, \rho]$ be independent random variables, and let $\zeta > 0$ be a sufficiently large absolute constant. Then for every $\varepsilon \in [0, 1]$, the summation $\sum_i Z_i$ is a $(\zeta\rho\frac{\log 1/\delta}{\varepsilon^2}, e^{\varepsilon})$-approximator to $\mathbb{E}[\sum_i Z_i]$, with probability $\geq 1 - \delta$.

Lemma 3.11 (Uniform Sampling): Fix $b \in \mathbb{N}$, $\varepsilon > 0$, and error probability $\delta > 0$. Consider some a_j, $j \in [b]$, such that $a_j \in [0, 1/b]$. For arbitrary $w \in [1, \infty)$, construct the set $J \subseteq [b]$ by subsampling each $j \in [b]$ with probability $p_w = \min(1, \frac{w}{b} \cdot \zeta\frac{\log 1/\delta}{\varepsilon^2})$. Then, with probability at least $1 - \delta$, the value $\frac{1}{p_w}\sum_{j \in J} a_j$ is a $(1/w, e^{\varepsilon})$-approximator to $\sum_{j \in [b]} a_j$, and $|J| \leq \mathcal{O}(w \cdot \frac{\log 1/\delta}{\varepsilon^2})$.

Lemma 3.12 (Non-uniform Sampling): Fix integers $n \leq N$, approximation $\varepsilon > 0$, factor $1 < f < 1.1$, error probability $\delta > 0$, and an "additive error bound" $\rho > 6n/\varepsilon/N^3$. There exists a distribution $\mathcal{W}$ on the real interval $[1, N^3]$ with $\mathbb{E}_{w \in \mathcal{W}}[w] \leq \mathcal{O}(\frac{1}{\rho} \cdot \frac{\log 1/\delta}{\varepsilon^3} \cdot \log N)$, as well as a "reconstruction algorithm" R, with the following property.

Take arbitrary $a_i \in [0, 1]$, for $i \in [n]$, and let $\sigma = \sum_i a_i$. Suppose one draws w_i i.i.d. from $\mathcal{W}$ and let $\hat{a}_i$ be an $(1/w_i, f)$-approximator of a_i. Then, given $\hat{a}_i$ and w_i for all $i \in [n]$, the algorithm R generates a $(\rho, f \cdot e^{\varepsilon})$-approximator to σ, with probability at least $1 - \delta$.