# Pattern Matching in Sub-Linear Time Using a Sparse Fourier Transform Approach

Nagaraj T. Janakiraman, Avinash Vem, Krishna R. Narayanan

Department of Electrical & Comp. Engg., Texas A&M University, College Station, TX, U.S.A

{tjnagaraj,vemavinash,krn}@tamu.edu

## Abstract

We consider the problem of querying a string (or, a database) of length $N$ bits to determine all the locations where a substring (query) of length $M$ appears either exactly or is within a Hamming distance of $K$ from the query. We assume that sketches of the original signal can be computed off line and stored. Using the sparse Fourier transform computation based approach introduced by Pawar and Ramchandran, we show that all such matches can be determined with high probability in sub-linear time. Specifically, if the number of matches is $L = N^\lambda$ and $M = N^\mu$, as $N \to \infty$, we show that the locations can be determined with a probability that approaches 1, with a computational complexity that is only $O\left(\max\left(\frac{N}{M}\log N, ML\log N\right)\right) = O\left(N^{\max(1-\mu, \mu+\lambda)}\log N\right)$ for $K \leq \frac{1}{6}M$. Further, the number of Fourier transform coefficients that need to be computed, stored and accessed, i.e., the sketching complexity of this algorithm is only $O\left(N^{\max(\mu, 1-\mu)}\right)$. Several extensions of the main theme are also discussed.

## I. Introduction and Problem Statement

We consider the pattern matching problem in the random setting which is a problem that has been studied extensively in computer science []. In this problem, we have a signal $\overrightarrow{x} := (x[1], x[2], \ldots, x[N])$ of length $N$ symbols representing a string, library or database. We first consider the random setting of the problem in which $x[i]$'s are a sequence of independent and identically distributed (i.i.d) random variables taking values in $\mathscr{A} := \{1, -1\}$, although extensions to other alphabets $\mathscr{A} \subseteq \mathbb{R}$ are possible. We assume that a sketch of $\overrightarrow{x}$ can be computed offline and stored, and the one-time computational complexity of creating the sketch is ignored.

**Exact Pattern Matching:** In the exact pattern matching problem, a substring of length $M$ of $\overrightarrow{x}$, namely $\overrightarrow{y} := \overrightarrow{x}[\tau : \tau + M - 1]$ is obtained by taking $M$ consecutive symbols from $\overrightarrow{x}$ and is presented as a query. This pattern may appear within $\overrightarrow{x}$ in $L$ different locations $\tau_1, \tau_2, \ldots, \tau_L$ and the objective is to determine all the locations $\tau_i, \forall i \in [1 : L]$. We consider the probabilistic version where our objective is to recover the matching locations with a probability that approaches 1 as $M, N \to \infty$.

**Approximate Pattern Matching:** In the approximate pattern matching version, $\overrightarrow{y}$ is a noisy version of a substring, i.e., $\overrightarrow{y} = \overrightarrow{x}[\tau : \tau + L - 1] \odot \overrightarrow{b}$ where $\overrightarrow{b}$ is a noise sequence with $b[i] \in \pm 1$ such that $d_H(\overrightarrow{y}, \overrightarrow{x}[\tau : \tau + M - 1]) \leq K$. Here $d_H$ denotes the Hamming distance, and $\odot$ represents component-wise multiplication. The objective is to determine all locations $\tau_i$ such that $d_H(\overrightarrow{y}, \overrightarrow{x}[\tau_i : \tau_i + M - 1]) \leq K$ with a probability that approaches 1 as $K, M$ and $N \to \infty$.

These problems are relevant to many applications including text matching, DNA sequencing, and is particularly relevant now due to the interest in applications involving huge volumes of data. The presented results are most useful in the following situations - (i) the string $\overrightarrow{x}$ is available before querying and one time computations such as computing a sketch of $\overrightarrow{x}$ can be performed off line and stored, and the complexity of computing the sketch of $\overrightarrow{x}$ can be ignored. Then, when queries in the form of $\overrightarrow{y}$ appear, one would like to decrease the computational complexity in searching for the string $\overrightarrow{y}$. (ii) The string $\overrightarrow{x}$ is not centrally available but parts of the string are sensed by different data collecting nodes distributively and communicated to a central server. A search query is presented to the server and the server needs to decide if the string appeared in the data sensed by one or more of the distributed nodes and the time instants where the query appeared. In this case, we would like to minimize the amount of data communicated by the nodes to the server and the computational complexity in searching for the

string. Finally, the proposed approach is most useful when the query $\overrightarrow{y}$ is not a pattern that can be predicted and, therefore, creating a look up table to quickly identify patterns is not efficient. For e.g., $\overrightarrow{x}$ is an i.i.d sequence of $\pm 1$ and $\overrightarrow{y}$ could be a substring of $\overrightarrow{x}$.

A naive approach to searching for a substring $\overrightarrow{y}$ in $\overrightarrow{x}$ is to compute the cross correlation between $\overrightarrow{x}$ and $\overrightarrow{y}$ denoted by $\underline{r} = [r[1], \ldots, r[N]]$

$$r[m] = (\overrightarrow{x} * \overrightarrow{y})[m] \triangleq \sum_{i=1}^{M} x[m+i-1]y[i] \tag{1}$$

and choosing the index $m$ that maximizes $r[m]$. This approach uses all the $N$ samples of $\overrightarrow{x}$ and has a super-linear computational complexity of $MN = N^{1+\mu}$. A computationally more efficient approach uses the fact that $\underline{r}$ can be computed by taking the inverse Fourier transform of the product of the Fourier transforms of $\overrightarrow{x}$ and $\overrightarrow{y}^*[-n]$, where $\overrightarrow{y}^*[-n]$ is the conjugate of time reversed $\overrightarrow{y}$. Such an approach still uses all the $N$ samples of $\overrightarrow{x}$ but reduces the computational complexity to $O(N \log N)$. Note that even though the Fourier transform of $\overrightarrow{x}$ can be precomputed, the $N$-point Fourier transform of $\overrightarrow{y}$ still needs to be computed resulting in the $O(N \log N)$ computational complexity.

Both the exact pattern matching problem and the approximate pattern matching problem have been extensively studied in computer science and two recent papers [1] and [2] provide a brief summary of the results. For the exact matching problem, Boyer and Moore presented an algorithm in [3] for finding the first occurrence of the match (only $\tau_1$) that has an expected complexity of $O(N/M \log N) = O(N^{1-\mu} \log N)$, whereas the worst case complexity (depending on $\tau_1$'s) can be $O(N \log N)$. For large $M$, the algorithm indeed has an average complexity that is sub-linear in $N$. This algorithm has been generalized to the approximate pattern matching problem in [4] with an average case complexity of $O(NK/M \log N)$ which provides a sub-linear time algorithm only when $K \ll M$. In the work by Andoni, Hassanieh, Indyk and Katabi [1], the authors gave the first sub-linear time algorithm with a complexity of $O(N/M^{0.359})$ even for $K = O(M)$.

## II. OUR MAIN RESULTS AND RELATION TO PRIOR WORK

Assuming that a sketch of $\overrightarrow{x}$ of size $O(N/M)$ can be computed and stored,

1) For the exact pattern matching problem, we present an algorithm which uses a sketching function for $\overrightarrow{y}$ with $O(\max(M \log N, N/M \log N))$ samples and a computational complexity of $O(\max(ML \log N, N/M \log N))$. When $M = N^{\mu}, L = N^{\lambda}$, this gives a sub-linear time algorithm.
2) For the approximate pattern matching problem, for $K = O(M)$, our algorithm has a sketching complexity of $\overrightarrow{y}$ is $O(\max(M \log N, N/M \log N))$ samples and computational complexity of $O(\max(ML \log N, N/M \log N))$ for all $K = O(M)$.

Our paper is inspired by and builds on two recent works by Hassanieh et al in [5] and Pawar and Ramchandran [6]. In [5], Hassanieh et al., considered the correlation function computation problem for a Global Positioning System (GPS) receiver and exploited the fact that the cross-correlation vector $\underline{r}$ is a very sparse signal in the time domain and, hence, the Fourier transform of $\underline{r}$ need not be evaluated at all the $N$ points. In the GPS application, which was the focus of [5], the query $\overrightarrow{y}$ corresponds to the received signal from the satellites and, hence, the length of the query was at least $N$. As a result, the computational complexity is still $O(N \log N)$ (still linear in $N$) and the only the constants were improved in relation to the approach of computing the entire Fourier transform. In a later paper by Andoni *et al.,* [1], a sub-linear time algorithm for shift finding in GPS is presented; however, this algorithm is completely combinatorial and eschews algebraic techniques such as FFT-based techniques.

There are important differences between our paper and [5], [1], [3], [2]. First and foremost, the algorithms used for pattern matching are entirely different. While their algorithms are combinatorial in nature, our algorithm is algebraic and uses signal processing and coding theoretic primitives. Secondly, the system model considered in our paper is different from the model in [5], [1], [3], [2] in that we allow for preprocessing or creating a sketch of the data $\overrightarrow{x}$. Our algorithm exploits this fact and results in a computational complexity $O(N/M)$ which is better than that in [1] for the approximate pattern matching problem. Finally, we also consider the problem of finding all matches of the pattern $\overrightarrow{y}$ instead of looking for only one match. In [6], Pawar and Ramchandran present an

algorithm based on aliasing and the peeling decoder for computing the Fourier transform of a signal with noisy observations for the case when the Fourier transform is known to be sparse. This algorithm has a complexity of $O(N \log N)$ and they do not consider the pattern matching problem. Our algorithm is similar to that of Pawar and Ramchandran's algorithm with one very important distinction. We modify their algorithm to exploit the fact that the peak of the correlation function of interest is always positive. This modification is key in computing the Sparse Inverse Discrete Fourier Transform (SIDFT) with sub-linear time complexity of $O(N^{1-\alpha} \log N)$, $0 < \alpha \leq 1$ . Thus, one of the main contributions of this paper is to show that signal processing and coding theoretic primitives, i.e., Pawar and Ramchandran's algorithm with some modifications can be used to solve the pattern matching algorithm in sub-linear time.

## III. NOTATIONS

This table below will introduce the notations we use in this paper.

| Symbol | Notational Meaning |
|---|---|
| $N$ | Size of the string or database in symbols |
| $M = N^\mu$ | Length of the query in symbols |
| $L = N^\lambda$ | Number of matches |
| $K$ | $\max_\tau d_H(\overrightarrow{x}[\tau : \tau + M - 1], \overrightarrow{y})$ |
| $G = N^\gamma$ | Number of blocks |
| $\tilde{N} = N^{1-\gamma}$ | Length of one block |
| $A = N^\alpha$ | Sub-sampling parameter |
| $B$ | Number of shifts |
| $d$ | Number of stages in the FFAST algorithm |

We denote vectors using the underbar, time domain signals using lowercase letters and the frequency domain signals using uppercase letter. For example $\overrightarrow{x} = (x[1], x[2], \cdots x[N])$ denotes a time domain signal with $i^{th}$ time component denoted by $x[i]$, and $\underline{X} = \mathscr{F}\overrightarrow{x}$ denotes the Fourier coefficients of $\overrightarrow{x}$. Matrices are denoted using uppercases letters. We differentiate a signal from a matrix by having a underbar for a signal vector. We denote the set $0, 1, 2 \cdots, N-1$ by $[N]$.

## IV. DESCRIPTION OF THE ALGORITHM

In this section we describe our algorithm, with sample and time complexities that are sub-linear in $N$, to find the locations $\underline{\tau} = (\tau_1, \tau_2, \cdots \tau_L)$ in the string $\overrightarrow{x}$, where the string (query) $\overrightarrow{y}$ matches. This is achieved by computing the cross-correlation $\underline{r}$ of $\overrightarrow{x}$ and $\overrightarrow{y}$, defined in Eqn. (1), and finding the positions where there is a significant peak.

The main idea here is that $\underline{R_{XY}}$ is sparse (upto some noise) with dominant peaks at $L$ positions ($\tau$) where the strings match, and noise components at $N - L$ positions where the strings do not match. Consider the case of exact matching,

$$r[m] = \begin{cases} M, & \text{if } m \in \mathscr{T} \\ n_m, & m \in [N] - \mathscr{T} \end{cases} \tag{2}$$

where, $\mathscr{T} := \{\tau_1, \tau_2, \cdots, \tau_L\}$, $n_m$ is the noise component that is induced due to correlation of two i.i.d. sequence of random variables each taking values from $\mathscr{A} := \{+1, -1\}$. The $\underline{r}$ can also be computed as shown below:

$$\underline{r} = \underbrace{\mathscr{F}_N^{-1}}_{\text{I}} \{ \underbrace{\mathscr{F}_N\{\overrightarrow{x}\}}_{\text{II}} \odot \underbrace{\mathscr{F}_N\{\overrightarrow{y'}\}}_{\text{III}} \} \tag{3}$$

where $\mathscr{F}_N\{\cdot\}$ and $\mathscr{F}_N^{-1}\{\cdot\}$ refers to $N$-point discrete Fourier transform and its inverse respectively, $\odot$ is the point-wise multiplication operation and $y'[n] = y^*[-n]$. Fig. 1 presents a notional diagram of our Algorithm.

As evident from Equation 3, our algorithm for computing $R_{XY}$ consists of three stages:
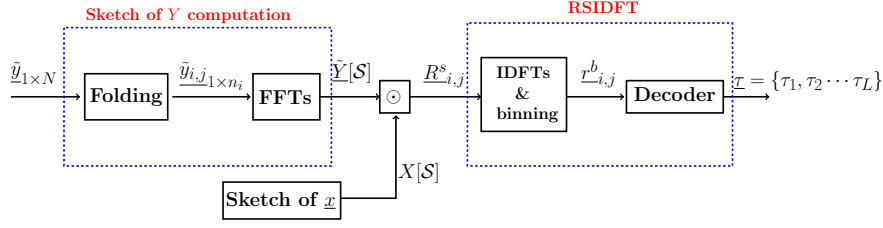
Fig. 1.

## I *Computing sparse $\mathscr{F}^{-1}$*:

Since $\underline{r}$ is sparse, we use a Robust Sparse Inverse Discrete Fourier Transform(RSIDFT) framework to compute the *L*-sparse coefficients. The architecture of RSIDFT is similar to FFAST proposed in [6], but the decoding algorithm has some key modifications to handle the noise model induced in this problem.

### A. RSIDFT Framework

Let $\underline{R} = \underline{X} \odot \underline{Y}'$ be the FFT of the cross-correlation of $\overrightarrow{x}$ and $\overrightarrow{y}$. We will see in Sec. IV-C how do we compute the required samples of $\underline{X}$ and $\underline{Y}$ but for this section we will assume that $\underline{X}$ and $\underline{Y}$ are available and the required samples of $\underline{R}$ is computed and input to RSIDFT framework. The RSIDFT framework computes the *L* dominant coefficients of $\underline{r}$ by using only a sub-sampled version of $\underline{R}$.

Consider the RSIDFT framework shown in Figure 2. The framework consists of *d-stages* with each stage correponding to a sub-sampling factor of $\frac{N}{f_i}$. The design scheme for choosing parameters $\{f_1, f_2, \ldots, f_d\}$ for various regimes of $\mu$ such that $f_i | N$ and $\tilde{f}_i = N^\alpha + O(1) \forall i$ are given in IV-B4. In each stage, there are *B branches* with shifts from $\underline{s} = [s_1, s_2, \cdots s_B]$, where $s_1 = 0$ in the first branch and the rest chosen randomly from $[N^\alpha]$. We can also carefully choose the shifts to satisfy mutual coherence property and restricted isometric property described in the analysis section.
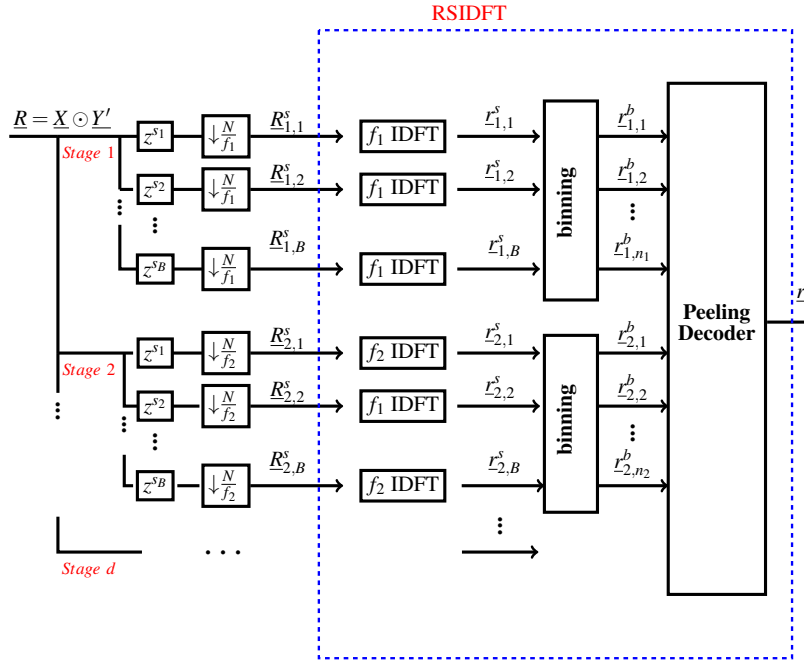


Fig. 2.  RSIDFT Framework to compute inverse Fourier Transform of a signal $\underline{R}$ that is sparse in time domain.

Given the input $\underline{R}$, in branch $j$ of $i^{\text{th}}$ stage or referred to as *branch* $(i,j)$ for convenience, RSIDFT sub-samples the signal $\underline{R}$ at

$$\mathscr{S}_{i,j} := \{s_j,\ s_j + n_i,\ s_j + 2n_i,\ \cdots s_j + (f_i - 1)n_i\}, \tag{4}$$

where $n_i := \frac{N}{f_i}$ to obtain $\underline{R}_{i,j}^s = \underline{R}[\mathscr{S}_{i,j}]$. The sub-sampling operation is followed by a $f_i$-point IDFT in each branch of stage $i$ to obtain $\underline{r}_{i,j}^s$. Notice that $\underline{r}_{i,j}^s$ is an aliased version of $\underline{r}$ due to the property that sub-sampling in Fourier domain is equivalent to aliasing in time domain.

Let $\underline{r}_{i,k}^b$ be the vector of observations that is formed by combining all $k^{\text{th}}$ coefficients of $\underline{r}_{i,k}^s$ (belonging to stage $i$) together as a vector, where $k \in [0 : f_i - 1]$, i.e.

$$\underline{r}_{i,k}^b = \begin{bmatrix} r_{i,1}^s[k] \\ r_{i,2}^s[k] \\ \vdots \\ r_{i,B}^s[k] \end{bmatrix}$$

More formally we can write the relationship between the the observation vectors $\underline{r}_{i,j}^b$ at bin $(i,j)$ and sparse vector to be estimated $\underline{r}$ as:

$$\underline{r}_{i,k}^b = \mathbf{W}_{i,k} \times \begin{bmatrix} r[k + (0)f_i] \\ r[k + (1)f_i] \\ \vdots \\ r[k + (n_i - 1)f_i] \end{bmatrix} \tag{5}$$

where we refer to $\mathbf{W}_{i,k}$ as the sensing matrix at bin $(i,k)$ and can be defined as

$$\mathbf{W}_{i,k} = \left[ \underline{w}^k, \underline{w}^{k+f_i}, \ldots, \underline{w}^{k+(n_i-1)f_i} \right] \text{ and } \underline{w}^k = \begin{bmatrix} e^{\frac{j2\pi k s_1}{N}} \\ e^{\frac{j2\pi k s_2}{N}} \\ \vdots \\ e^{\frac{j2\pi k s_B}{N}} \end{bmatrix} \tag{6}$$



$$w_i^k := \left( e^{j\frac{2\pi s_i}{N}} \right)^k$$

$$\underline{r}_{1,1}^b = \begin{bmatrix} r_{1,1}^s[1] = r[0]w_1^0 + r[3]w_1^3 \\ r_{1,2}^s[1] = r[0]w_2^0 + r[3]w_2^3 \end{bmatrix}$$

$$\underline{r}_{1,2}^b = \begin{bmatrix} r_{1,1}^s[2] = r[1]w_1^1 + r[4]w_1^4 \\ r_{1,2}^s[2] = r[1]w_2^1 + r[4]w_2^4 \end{bmatrix}$$

$$\underline{r}_{1,3}^b = \begin{bmatrix} r_{1,1}^s[3] = r[2]w_1^2 + r[5]w_1^5 \\ r_{1,2}^s[3] = r[2]w_2^2 + r[5]w_2^5 \end{bmatrix}$$

$$\underline{r}_{2,1}^b = \begin{bmatrix} r_{2,1}^s[1] = r[0]w_1^0 + r[2]w_1^2 + r[4]w_1^4 \\ r_{2,2}^s[1] = r[0]w_2^0 + r[2]w_2^2 + r[4]w_2^4 \end{bmatrix}$$

$$\underline{r}_{2,2}^b = \begin{bmatrix} r_{2,1}^s[2] = r[1]w_1^1 + r[3]w_1^3 + r[5]w_1^5 \\ r_{2,2}^s[2] = r[1]w_2^1 + r[3]w_2^3 + r[5]w_2^5 \end{bmatrix}$$
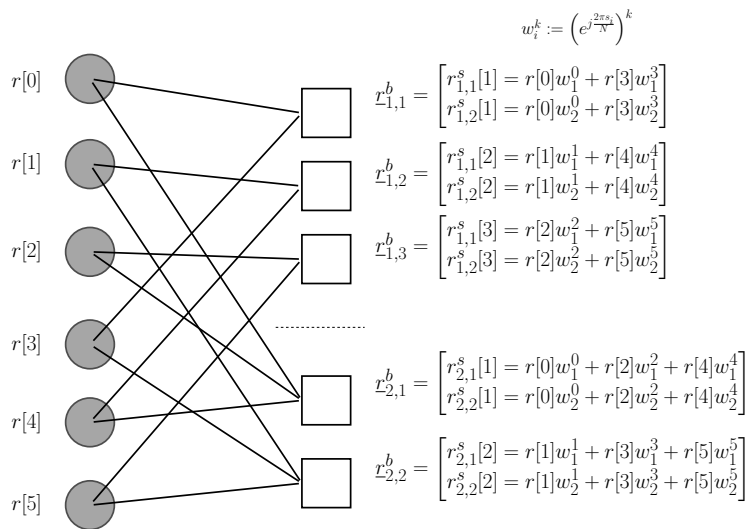
Fig. 3. Example of a Tanner graph formed in a RSIDFT framework with system parameters being $d = 2$, $B = 2$, $N = 6$, $f_1 = 2$ and $f_2 = 3$. The variable nodes (colored gray circles) represent the cross-correlation vector $\underline{r}$ and the bin nodes (uncolored white boxes) represent the binned observation vector $\underline{r}_{i,k}^b$. The figure also illustrates the relationship between $\underline{r}_{i,k}^b$ and $\underline{r}$.

In Fig. 3 we represent the relation between $\underline{r}$ and $\tilde{\underline{r}}^b := [\underline{r}^b_{1,1}, \underline{r}^b_{1,2}, \ldots, \underline{r}^b_{1,n_1}, \ldots, \underline{r}^b_{2,1}, \ldots, \underline{r}^b_{2,n_2}, \ldots, \underline{r}^b_{B,1}, \ldots, \underline{r}^b_{B,n_B}]$ via an example using Tanner graph. We refer to $\tilde{\underline{r}}^b$ as sensing signal. The nodes on the left, which we refer to as *variable nodes*, represent the $N$ elements of vector $\underline{r}$. And similarly the nodes on the right, which we refer to as *bin nodes*, represent the $\sum_{i \leq B} n_i$ sub-sensing signals. We will now describe the decoding algorithm which takes the sensing signal $\tilde{\underline{r}}^b$ as input and estimates the $L$-sparse $\underline{r}$.

### B. Decoder

Observe from the Tanner graph that the degree of each bit node is $d$ and that of each bin node is approximately $N^\alpha$. We refer to a bin node as *zero-ton* (or $\mathcal{H}_z$) if the number of variable nodes with significant amplitude connected to the bin node is zero. The *singleton ($\mathcal{H}_s$), double-ton ($\mathcal{H}_d$) and multi-ton ($\mathcal{H}_m$)* bin nodes are defined similarly where the number of significant variable nodes connected are one, two and greater than two, respectively. The peeling decoder has the following three steps in the decoding process.

*1) Bin Classification:* In this step a bin node is classified as a zero-ton or a singleton or a multi-ton. At bin $(i,j)$ the classification is done based on comparing the first observation $r^b_{i,j}[1]$, which corresponds to zero shift, with a predefined threshold. Let the value of $r^b_{i,j}[1]$ be $z$, then the classification rules can be written as follows:

$$\widehat{\mathcal{H}}_{i,j} = \begin{cases} \mathcal{H}_z & z/M < \gamma_1 \\ \mathcal{H}_s & \gamma_1 < z/M < \gamma_2 \\ \mathcal{H}_d & \gamma_2 < z/M < \gamma_3 \\ \mathcal{H}_m & z/M > \gamma_3 \end{cases} \tag{7}$$

where $(\gamma_1, \gamma_2, \gamma_3) = (\frac{1-2\eta}{2}, \frac{3-4\eta}{2}, \frac{5-6\eta}{2})$. Note that for the case of exact matching $\eta = 0$.

*2) Singleton decoding:* If a bin node $(i,j)$ is classified as a singleton, we need to identify the position of the significant non-zero variable node connected to it. This is done by correlating the observation vector $\underline{r}^b_{i,j}$ with each column of $\mathbf{W}_{i,j} := [\underline{w}^j, \underline{w}^{j+n_i}, \ldots, \underline{w}^{j+(f_i-1)n_i}]$ and choose the index that maximizes the correlation value.

$$\hat{k} = \arg \max_{k \in \{j+ln_i\}} \underline{w}^k \, \underline{r}^b_{i,j}$$

The value of the variable node connected to the singleton bin is estimated as follows:

$$\hat{r}[\hat{k}] = M(1-\eta).$$

Note that for the case of exact matching we know the value to be exactly equal to $M$ which is our estimate in the case of exact matching ($\eta = 0$). But in the case of approximate matching the actual value of $x[k] \in \{M(1-2\eta), \ldots, M-1, M\}$ and our estimate $\hat{r}[\hat{k}] = M(1-\eta)$, the mid-point of the range, is only a loose approximate, this would suffice for our purposes as we are mainly interested in recovering only the positions of matches i.e. the indices of the sparse coefficients in $\underline{r}$ nit their values.

*3) Peeling Process:* The main idea behind the peeling based decoder is to find a singleton bin, identify the significant variable node connected to the bin, decode it's value and remove it's contribution from all the bin nodes connected to that variable node. Although the main idea is similar for exact matching and the approximate matching scenarios, there are some subtle differences in their implementation.

*Exact Matching*: In the case of exact matching, we remove the decoded variable node's contribution from all the bin nodes it is connected to.
*Approximate Matching*: In this case we remove the decoded variable node's contribution only from singleton and double-ton bins and do not alter the bins which are classified to be multi-tons with degree more than two.

We present the overall recovery algorithm, which comprises of *bin classification, singleton decoding* and *peeling process*, in the Algorithm 1 pseudo-code.

---

**Algorithm 1** Peeling based recovery algorithm

---

Compute $\widehat{\mathcal{H}}_{i,j}$ for $i \in [B], j \in [n_i]$. See Eqn. (7)

**while** $\exists\, i, j : \widehat{\mathcal{H}}_{i,j} = \mathcal{H}_s$, **do**

    $(\hat{k}, \hat{r}[\hat{k}]) =$**Singleton-Decoder**$(\underline{\mathrm{r}}^b_{i,j})$

    Assign $\hat{r}[\hat{k}]$ to $\hat{k}^{\text{th}}$ variable in bin $(i, j)$

    **for** $(i_0, j_0) \in \mathcal{N}(\hat{k})$ **do**

        $\underline{\mathrm{r}}^b_{i_0,j_0} \leftarrow \underline{\mathrm{r}}^b_{i_0,j_0} - \hat{r}[\hat{k}]\mathbf{g}_{\hat{k}}$                                             *Exact Matching*

        $\underline{\mathrm{r}}^b_{i_0,j_0} \leftarrow \underline{\mathrm{r}}^b_{i_0,j_0} - \hat{r}[\hat{k}]\mathbf{g}_{\hat{k}}$      only if $\widehat{\mathcal{H}}_{i_0,j_0} = \mathcal{H}_s$ or $\mathcal{H}_d$           *Approximate Matching*

        Recompute $\widehat{\mathcal{H}}_{i_0,j_0}$

---

**Algorithm 2** Singleton-Decoder

---

**Input:** $\underline{\mathrm{r}}^b_{i,j}$

**Output:** $(\hat{k}, \hat{r}[\hat{k}])$

Estimate singleton index to be $\hat{k} = \underset{k \in \{j+ln_i\}}{\arg\max}\ \underline{w}^k\ \underline{\mathrm{r}}^b_{i,j}$

Estimate the value to be:

$$\hat{r}[\hat{k}] = \begin{cases} M & \text{Exact Matching case} \\ M - \frac{K}{2} & \text{Approximate Matching case} \end{cases}$$

---

*4) Choosing $\alpha$ and $f_i$ for various $\mu$:* For a given value of $\mu$, we will describe how to choose the parameters $d$ and $f_i$.

*Exact Matching*: Find $d \in \mathbb{Z}_+$ such that $\mu \in (\frac{1}{d}, \frac{1}{d-1}) \cup (1 - \frac{1}{d-1}, 1 - \frac{1}{d})$

*Approximate Matching*: If $\mu \in (\frac{1}{8}, \frac{7}{8})$, choose $d = 8$. Else find $d \in \mathbb{Z}_+$ such that $\mu \in (\frac{1}{d}, \frac{1}{d-1}) \cup (1 - \frac{1}{d-1}, 1 - \frac{1}{d})$ and $d \geq 8$

- Find a factorization for signal length $N = \prod_{i=0}^{d-1} \mathcal{P}_i$ such that the set of integers $\{\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_{d-1}\}$ are pairwise co-prime and all the $\mathcal{P}_i$ are approximately equal
- More precisely, let $\mathcal{P}_i = \mathbf{F} + O(1)\ \forall i$ and choose $f_i = N/\mathcal{P}_i$
- Our choice of design parameters results in $f_i \approx N^\alpha$ where $\alpha = 1 - \frac{1}{d}$.

To summarize our choice of design parameters, both in the exact and approximate matching cases, for any $0 < \mu < 1$, we choose the down-sampling factors $f_i$ to be approximately equal to $N^\alpha$ in all the $d$-stages where $\alpha$ satisfies:

$$\alpha > \max(\mu, 1 - \mu) \tag{8}$$

which is easy to verify.

*C. Sketches of $\underline{X}$ and $\underline{Y}$*

As we have already seen in Sec. IV-A the RSDIFT framework requires the values of $\underline{R}(= \underline{X} \odot \underline{Y})$ only at indices $\mathscr{S}$ or in other words we need $\underline{X}$ and $\underline{Y}$ only at the indices $\mathscr{S}$.

*Computing the sketch of $\overrightarrow{\mathrm{x}}$*: We assume that the sketch of $\overrightarrow{\mathrm{x}}$, $\underline{X}[\mathscr{S}] = \{X[i], i \in \mathscr{S}\}$ is pre-computed and stored in a database.

*Computing the sketch of $\overrightarrow{\mathrm{y}}$*: For every new query $\overrightarrow{\mathrm{y}}$, only $\{\underline{Y}'[\mathscr{S}_{i,j}], i \in [d], j \in [B]\}$ needs to be computed where the subsets $\mathscr{S}_{i,j}$ are defined in Eq. (4). Naively, the FFT algorithm can be used to compute $\underline{Y}'$ and the required subset of coefficients can be taken but this would be of $O(N \log N)$ complexity. We observe that $\underline{Y}'[\mathscr{S}_{i,j}]$ is $\underline{Y}'$ shifted by $s_j$ and sub-sampled by a factor of $f_i$. Thus for a given $(i, j)$ this corresponds to, in time

domain, a point-wise multiplication by $[1, W_{s_j}, W_{s_j}^2, \ldots, W_{s_j}^{N-1}]$ followed by *folding* the signal into $f_i$ number of length $\frac{N}{f_i}$ signals and adding them up resulting in a single length $\frac{N}{f_i}$ signal denoted by $\overrightarrow{\mathbf{y}}'_{i,j}$. Formally the *folding* operation can be described as follows:

$$\overrightarrow{\mathbf{y}}'_{i,j} = \sum_{m=0}^{f_i-1} \overrightarrow{\mathbf{y}}'[mn_i : (m+1)n_i - 1] \odot [W_{s_j}^{mn_i}, W_{s_j}^{mn_i+1}, \ldots, W_{s_j}^{(m+1)n_i-1}], \qquad W_{s_j} = e^{\frac{j2\pi s_j}{N}} \qquad (9)$$

where $n_i = \frac{N}{f_i}$. Taking $n_i$-point DFT of $\overrightarrow{\mathbf{y}}'_{i,j}$ produces $\underline{Y}'[\mathscr{S}_{i,j}]$ i.e. $\underline{Y}'$ sampled only at $\mathscr{S}_{i,j}$ indices which is what we need in branch $(i, j)$. To obtain all the samples in $\mathscr{S}$ required for the RSDIFT framework, the *folding* technique needs to be carried out for each $(i, j)$, for $i \in [d], j \in [B]$, a total of $\approx dB \ N^{1-\alpha}$-point DFTs.

# V. Performance Analysis

In this section, we will analyze the overall probability of error involved in finding the correct matching positions. This can be done by analyzing the following three error events independently and then using a union bound to bound the total probability of error.

- $\mathscr{E}_1$-*Bin Classification*: Event that a bin is wrongly classified
- $\mathscr{E}_2$-*Position Identification*: Given a bin is correctly identified as a singleton, event that the position of singleton is identified incorrectly
- $\mathscr{E}_3$-*Peeling Process*: Given the classification of all the bins and the position identification of singletons in each iteration is accurate, event that the peeling process fails to recover the $L$ significant correlation coefficients

## A. Bin Classification

**Lemma 1.** *The probability of bin classification error at any bin $(i, j)$ can be upper bounded by*

$$\mathbb{P}[\mathscr{E}_1] \leq 6e^{-\frac{N^{\mu-\alpha}(1-6\eta)^2}{16}}$$

*Proof.*

$$\mathbb{P}[\mathscr{E}_1] = \mathbb{P}[\mathcal{H}_z]\mathbb{P}[\mathscr{E}_1|\widehat{\mathcal{H}}_{i,j} = \mathcal{H}_z] + \mathbb{P}[\mathcal{H}_s]\mathbb{P}[\mathscr{E}_1|\widehat{\mathcal{H}}_{i,j} = \mathcal{H}_s] + \mathbb{P}[\mathcal{H}_d]\mathbb{P}[\mathscr{E}_1|\widehat{\mathcal{H}}_{i,j} = \mathcal{H}_d \cup \mathcal{H}_m]$$

$$\leq \mathbb{P}[\mathscr{E}_1|\widehat{\mathcal{H}}_{i,j} = \mathcal{H}_z] + \mathbb{P}[\mathscr{E}_1|\widehat{\mathcal{H}}_{i,j} = \mathcal{H}_s] + \mathbb{P}[\mathscr{E}_1|\widehat{\mathcal{H}}_{i,j} = \mathcal{H}_d \cup \mathcal{H}_m]$$

$$\leq e^{-\frac{N^{\mu-\alpha}(1-2\eta)^2}{8}} + 2e^{-\frac{N^{\mu-\alpha}(1-4\eta)^2}{16}} + 2e^{-\frac{N^{\mu-\alpha}(1-6\eta)^2}{16}} + e^{-\frac{N^{\mu-\alpha}(1-6\eta)^2}{16}}$$

$$\leq 6e^{-\frac{N^{\mu-\alpha}(1-6\eta)^2}{16}}$$

where the inequalities in the third line are due to Lemmas 9, 10, 11 and 12 respectively provided in Appendix B. $\qquad \square$

## B. Position Identification

**Lemma 2.** *Given that a bin $(i, j)$ is correctly classified as a singleton, the probability of error in identifying the position of the non-zero variable node can be upper bounded by*

$$\mathbb{P}[\mathscr{E}_2|\overline{\mathscr{E}}_1] \leq e^{-\frac{N^{\mu-\alpha}(1-2\eta)^2}{8}} + e^{-\frac{N^{\mu-\alpha}(1-2\eta)^2}{8}}$$

*C.* **Peeling Process**

To analyze the peeling process alone independently, we refer to a *oracle based peeling decoder* which has the accurate classification of all the bins and given a singleton bin, at any iteration, can accurately identify the position of the singleton. In other words, oracle based peeling decoder is the peeling part of our decoding scheme but with the assumption that the bin classification and position identification are carried out without any error.

**Lemma 3** (Exact Matching). *The oracle based peeling decoder in the exact matching case, for $d = 3$ and $F = \eta L$ for any $\eta > 0.4073$:*

- *successfully uncovers all the matching positions with probability at least $1 - O(1/dF)$*
- *successfully uncovers all but a vanishingly small fraction, i.e. $o(L)$, of the matching positions with probability at least $1 - e^{-\beta \varepsilon_1^2 k^{1/(4l+1)}}$ for some constants $\beta, \varepsilon_1 > 0$ and $l > 0$.*

*Proof.* We borrow this result from Pawar and Ramchandran's [6]. Although our RSDIFT framework and their robust-FFAST scheme have two main differences:

- We are computing smaller IDFT's to recover a sparse bigger IDFT whereas in [6] the same can be said for DFT instead of IDFT
- The second major difference is in the decoding scheme: Our problem model is such that the sparse components of the signal space has only positive amplitude and thus our bin processing part (bin classification and position identification) is different than in [6]

Irrespective of these two differences, the Tanner graph representation of the framework and the peeling part of the decoder are identical to that of the robust-FFAST scheme. And thus the limits of the *oracle based peeling decoder* for our scheme, which assumes the bin processing is accurate thus making the analysis independent of bin decoder, will be identical to that of the peeling part of the decoding scheme in robust-FFAST. □

In any iteration, given a singleton bin, the peeling process, in the case of approximate matching, runs the Singleton-Decoder algorithm on the bin only if it was either originally a singleton or originally a double-ton with one of the variable nodes being peeled off already. Whereas in the case of exact matching case, the peeling decoder runs the Singleton-Decoder on the bin irrespective of it's original degree. Hence we need to analyze the oracle based peeling decoder for the approximate matching case differently compared to the exact matching case.

**Lemma 4** (Approximate Matching). *The oracle based peeling decoder in the approximate matching case: for $d = 8$ and $F = \eta L$ where $\eta > 0.76625$:*

- *successfully fails to uncover all but a small fraction $\varepsilon = 10^{-3}$ of the matching positions with exponentially decaying probability.*

*For higher target $\varepsilon$, we need to give the higher values of $d$ and $\eta$ according to Group testing paper. But the difficulty is in proving that for which we need the equivalence of the CRT construction with random construction and hence requires careful factorization of $f_1, f_2, \ldots$.*

# VI. COMPLEXITY ANALYSIS

In this section, we will analyze the number of samples used from the actual database to run the algorithm and also the computational complexity as a function of the system parameters.

*A. Sample Complexity*

As stated in the problem statement, we exploit the sparseness in IDFT computations to compute the correlations using less number of samples from the database.

In each branch of the FFAST algorithm, executed over a block of samples, we down-sample the $\tilde{N}$ samples by a factor of $N^{\alpha}$ to get $N^{1-\gamma-\alpha}$ samples. We also do this for each chosen shift and we get a total of $N^{1-\gamma-\alpha+\beta}$

samples per block per stage. We repeat this for $d = \frac{1-\gamma}{\alpha}$ such stages to get a total complexity of $dN^{1-\gamma-\alpha+\beta}$ samples per block. So, the total number of samples is given by

$$\text{Total \# of samples required (S)} = \underbrace{\frac{1-\gamma}{\alpha}}_{d \text{ stages}} \underbrace{N^{1-\gamma-\alpha+\beta}}_{\text{Samples per block per stage}} \underbrace{N^{\gamma}}_{\text{\# of blocks}} = \frac{1-\gamma}{\alpha} N^{1-\alpha+\beta}$$

## B. Computational Complexity

Consider the operations involved in the computation of correlation of $X$ and $Y$, $R_{XY}$, as shown below.

$$R_{XY} = \underbrace{IDFT}_{I} \underbrace{\{DFT\{X\}}_{II} \times \underbrace{DFT\{Y\}\}}_{III}$$

In each block of samples from the database, the process of computing the correlation involves three sets of operations: two $\tilde{N}$-point DFT operations ($II, III$) sampled at $N^{1-\gamma-\alpha+\beta}$ points(based on sampling requirements of FFAST architecture) and one sparse-IDFFT operation ($I$). We assume that the database already contains sampled version of DFT $\{ X \}$ stored and hence we exclude those computations from the complexity calculations. So, it is enough that we analyze the computations involved in operations ($I, III$) on each block to get the overall computational complexity.

1) Operation - III:
Notice that the operation $DFT\{Y\}$ is just a one time computation and can be reused in each block since the query is same for all blocks. Since we only need to compute the sampled version of the DFT$\{Y\}$ in each branch, we can do some clever implementations to reduce the complexity. We can create aliasing on $Y$ and then take a shorter length ($N^{1-\gamma-\alpha}$-point) DFT to obtain the sampled version of $\tilde{N}$-point DFT of $Y$.
This method of computing sampled DFT$\{Y\}$ involves two operations - folding and adding (aliasing), and then a $N^{1-\gamma-\alpha}$-point DFT. The complexity involved in computing down-sampled DFT$\{Y\}$, $C_{III}$, is given by

$$C_{III} = \underbrace{\frac{1-\gamma}{\alpha}}_{d \text{ stages}} ( \underbrace{N^{\delta+\beta}}_{\text{\# of additions (aliasing)}} + \underbrace{N^{1-\gamma-\alpha+\beta} \log N^{1-\gamma-\alpha}}_{\text{Shorter FFTs}} )$$

Folding and adding, for each shift, involves adding $N^{\alpha}$ vectors of length $N^{1-\gamma-\alpha}$. We know that the length of the query is $L = N^{\delta}$, i.e., the number of non-zero elements in $Y$ (zero-padded version of the query) is $L$ and hence we only need to compute $L$ additions instead of the length of the vector.

2) Operation - I:
Taking advantage of the fact that the correlations are sparse (with some noise), we use a FFAST based architecture to reduce the complexity involved in computing IDFT. The number of computations involved in this process, $C_I$, is given by

$$C_I = \underbrace{N^{\gamma}}_{\text{\# of blocks}} \underbrace{\frac{1-\gamma}{\alpha}}_{d \text{ stages}} \underbrace{N^{1-\gamma-\alpha+\beta} \log N^{1-\gamma-\alpha}}_{\text{Shorter IFFTs /block/stage}} = \frac{1-\gamma}{\alpha} N^{1-\alpha+\beta} \log N^{1-\gamma-\alpha}$$

Total number of computations, $C = C_{III} + C_I$, is given by

$$C = \frac{1-\gamma}{\alpha} (N^{\delta+\beta} + N^{1-\gamma-\alpha+\beta} \log N^{1-\gamma-\alpha} + N^{1-\alpha+\beta} \log N^{1-\gamma-\alpha})$$

**Lemma 5** (Chernoff Bound for bounded random variables). *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables such that $X_i \in \{-1, +1\}$ for all $i$ and $E[X_i] = \mu$ for all $i$. Then for any $\delta > 0$:*

$$\textbf{\textit{Upper Tail}} : \mathbb{P}\left[\frac{1}{n}\sum (X_i - \mu) \geq \delta\right] \leq e^{-\frac{n\delta^2}{2}}$$

$$\textbf{\textit{Lower Tail}} : \mathbb{P}\left[\frac{1}{n}\sum (X_i - \mu) \leq -\delta\right] \leq e^{-\frac{n\delta^2}{4}}$$

**Lemma 6** (Variant of Hoeffding Bound). *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables such that $X_i \in [a_i, b_i]$ and $E[X_i] = \mu$ for all $i$. Then for any $\delta > 0$:*

$$\textbf{\textit{Upper Tail}} : \mathbb{P}\left[\sum (X_i - \mu) \geq \delta\right] \leq \exp\left\{-\frac{2\delta^2}{\sum (b_i - a_i)^2}\right\}$$

**Lemma 7.** *Let us consider $r[\theta_0], r[\theta_0 + f_i], \ldots, r[\theta_0 + (n_i - 1)f_i]$ where $\theta_0 + kf_i \notin \{\tau_1, \ldots, \tau_L\}$ are not one of the matching positions. Then for any $\delta > 0$:*

$$\textbf{\textit{Upper Tail}} : \mathbb{P}\left[\left(\sum_{i=0}^{n_i-1}\sum_{k=0}^{M-1} x[\theta_i + k]y[k]\right) \geq M\delta\right] \leq \exp\left\{-\frac{M\delta^2}{2n_i}\right\}$$

*Proof.* Then $\mathbb{P}[x[\theta_i + k]y[k] = +1]$ with probability $1/2$ and $\mathbb{E}[x[\theta_i + k]y[k]] = 0$. We will use the fact that $\alpha > \mu$ from Eqn. (8) and thus $f_i > M$ and hence these terms are non-overlapping. Thus we can show the independence and hence we can apply the bounds from Lem. 6. $\qquad\square$

**Remark 8.** *Let us consider $r[\theta_1], \ldots, r[\theta_{f_i}]$ where $\theta_i \notin \{\tau_1, \ldots, \tau_L\}$ are not one of the matching positions. Then we can show that $\mathbb{P}[x[\theta_i + k]y[k] = +1]$ with probability $1/2$ and $\mathbb{E}[x[\theta_i + k]y[k]] = 0$. We also need to show that the set of random variables $\{x[\theta_i + k]y[k] : i \in \{1, 2, \ldots f_i\}, k \in [M]\}$ are independent to be able to apply Chernoff bounds or Normal distribution approximations. I was able to show this for the case of $M = 3$. I don't see a simple way of extending this to the case of general $M$. Note that without independence of these terms from non-matching positions, we can only obtain polynomial decay from Chebyshev inequality that requires only pairwise independence.*

We employ classification rules based only on the first element of the measurement vector at bin $(i, j)$ which can be given by

$$Z[1] = \begin{cases} \sum_{\ell=0}^{f_i-1}\sum_{k=0}^{M-1} n_{l,k} & \text{if} \quad \mathcal{H} = \mathcal{H}_z \\ M_1 + \sum_{\ell=0}^{f_i-2}\sum_{k=0}^{M-1} n_{l,k} & \text{if} \quad \mathcal{H} = \mathcal{H}_s \\ M_1 + M_2 + \sum_{\ell=0}^{f_i-3}\sum_{k=0}^{M-1} n_{l,k} & \text{if} \quad \mathcal{H} = \mathcal{H}_d \end{cases} \tag{10}$$

where $n_{l,k} = x[\theta_\ell + k]y[k]$ and $\theta_\ell \notin \{\tau_1, \tau_2, \ldots, \tau_L\}$. Also for the case of exact matching $M_1 = M_2 = M$ whereas in the case of approximate matching the values of $M_1, M_2 \in [M(1 - 2\eta) : M]$.

**Lemma 9** (zero-ton). *Given that the bin $(i, j)$ is a zero-ton, the classification error can be bounded by*

$$\mathbb{P}[\mathcal{E}_1 | \mathcal{H}_z] \leq e^{-\frac{N^{\mu-\alpha}(1-2\eta)^2}{8}}$$

*Proof.* The above expression can be derived by observing that a bin is not classified as zero-ton if $\frac{Z[1]}{M} \geq \frac{1-2\eta}{2}$. Let us denote the probability of this event as $p_{z1}$ which can be bounded as:

$$
\begin{aligned}
p_{z1} &= \mathbb{P}\left[\frac{Z[1]}{Mf_i} \geq \frac{1-2\eta}{2f_i}\right] \\
&\leq e^{-\frac{Mf_i(1-2\eta)^2}{8f_i^2}} \\
&\approx e^{-\frac{N^{\mu-\alpha}(1-2\eta)^2}{8}}
\end{aligned}
$$

where the second bound is due to Eqn. (10) and Lemma. 5. The approximation in the third line is from our design that all the $f_i$ are chosen such that $f_i \approx N^\alpha$ and $M = N^\mu$. □

**Lemma 10** (singleton). *Given that the bin $(i,j)$ is a singleton, the classification error can be bounded by*

$$
\mathbb{P}[\mathcal{E}_1|\mathcal{H}_s] \leq 2e^{-\frac{N^{\mu-\alpha}(1-4\eta)^2}{16}}
$$

*Proof.* We observe that a bin is not classified as singleton if $\frac{Z[1]}{M} \leq \frac{1-2\eta}{2}$ or $\frac{Z[1]}{M} \geq \frac{3-4\eta}{2}$. Let us denote the probability of the two events as $p_{s1}$ and $p_{s2}$ respectively which can be bounded as:

$$
\begin{aligned}
p_{s1} &= \mathbb{P}\left[\frac{1}{M}\sum_{\ell=0}^{f_i-2}\sum_{k=0}^{M-1} n_{l,k} \leq \frac{1-2\eta}{2} - \frac{M_1}{M}\right] \\
&\leq \mathbb{P}\left[\frac{1}{M}\sum_{\ell=0}^{f_i-2}\sum_{k=0}^{M-1} n_{l,k} \leq -\frac{1-2\eta}{2}\right] \\
&\leq e^{-\frac{Mf_i(1-2\eta)^2}{16f_i^2}} \\
&\approx e^{-\frac{N^{\mu-\alpha}(1-2\eta)^2}{16}}
\end{aligned}
$$

where we used *lower tail* of Lemma 5 and $f_i \approx N^\alpha$ and the lower bound on $\frac{M_1}{M} \geq (1-2\eta)$. Similarly $p_{s2}$ can be upper bounded by:

$$
\begin{aligned}
p_{s2} &= \mathbb{P}\left[\frac{1}{M}\sum_{\ell=0}^{f_i-2}\sum_{k=0}^{M-1} n_{l,k} \geq \frac{3-4\eta}{2} - \frac{M_1}{M}\right] \\
&\leq \mathbb{P}\left[\frac{1}{Mf_i}\sum_{\ell=0}^{f_i-2}\sum_{k=0}^{M-1} n_{l,k} \geq -\frac{1-4\eta}{2f_i}\right] \\
&\approx e^{-\frac{N^{\mu-\alpha}(1-4\eta)^2}{8}}
\end{aligned}
$$

Thus the overall probability of error for classifying a singleton can be obtained by combining $p_{s1}$ and $p_{s2}$. □

**Lemma 11** (double-ton). *Given that the bin $(i,j)$ is a double-ton, the classification error can be bounded by*

$$
\mathbb{P}[\mathcal{E}_1|\mathcal{H}_d] \leq 2e^{-\frac{N^{\mu-\alpha}(1-6\eta)^2}{16}}
$$

*Proof.* We observe that a bin is not classified as double-ton if $\frac{Z[1]}{M} \leq \frac{3-4\eta}{2}$ or $\frac{Z[1]}{M} \geq \frac{5-6\eta}{2}$. Let us denote the probability of these two events as $p_{d1}$ and $p_{d2}$ respectively which can be bounded similar to Lemma 10.

$$
\begin{aligned}
p_{d1} &= \mathbb{P}\left[\frac{1}{M}\sum_{\ell=0}^{f_i-3}\sum_{k=0}^{M-1} n_{l,k} \leq \frac{3-4\eta}{2} - \frac{M_1+M_2}{M}\right] \\
&\leq \mathbb{P}\left[\frac{1}{M}\sum_{\ell=0}^{f_i-3}\sum_{k=0}^{M-1} n_{l,k} \leq -\frac{1-4\eta}{2}\right] \\
&\leq e^{-\frac{M(f_i-2)(1-4\eta)^2}{16(f_i-2)^2}} \\
&\approx e^{-\frac{N^{\mu-\alpha}(1-4\eta)^2}{16}}.
\end{aligned}
$$

Similarly $p_{d2}$ can be bounded as

$$
\begin{aligned}
p_{d2} &= \mathbb{P}\left[\frac{1}{M}\sum_{\ell=0}^{f_i-3}\sum_{k=0}^{M-1} n_{l,k} \geq \frac{5-6\eta}{2} - \frac{M_1+M_2}{M}\right] \\
&\leq \mathbb{P}\left[\frac{1}{M}\sum_{\ell=0}^{f_i-3}\sum_{k=0}^{M-1} n_{l,k} \geq \frac{1-6\eta}{2}\right] \\
&\leq e^{-\frac{M(f_i-2)(1-6\eta)^2}{8(f_i-2)^2}} \\
&\approx e^{-\frac{N^{\mu-\alpha}(1-6\eta)^2}{8}}
\end{aligned}
$$

where we use the lower bounds $M_1, M_2 \leq M$. $\qquad \square$

**Lemma 12** (multi-ton)**.** *Given that the bin $(i,j)$ is a multi-ton, the classification error can be bounded by*

$$
\mathbb{P}[\mathscr{E}_1|\mathcal{H}_m] \leq e^{-\frac{N^{\mu-\alpha}(1-6\eta)^2}{16}}
$$

*Proof.* We observe that a bin is not classified as multi-ton if $\frac{Z[1]}{M} \leq \frac{5-6\eta}{2}$. Let us denote the probability of this event as $p_{m1}$ which can be bounded as:

$$
\begin{aligned}
p_{m1} &= \mathbb{P}\left[\frac{1}{M}\sum_{\ell=0}^{f_i-3}\sum_{k=0}^{M-1} n_{l,k} \leq \frac{5-6\eta}{2} - \frac{M_m}{M}\right] \\
&\leq \mathbb{P}\left[\frac{1}{M}\sum_{\ell=0}^{f_i-m}\sum_{k=0}^{M-1} n_{l,k} \leq -\frac{1-6\eta}{2}\right] \\
&\leq e^{-\frac{M(f_i-m)(1-4\eta)^2}{16(f_i-m)^2}} \\
&\leq e^{-\frac{M(1-6\eta)^2}{16f_i}} \\
&\approx e^{-\frac{N^{\mu-\alpha}(1-6\eta)^2}{16}}.
\end{aligned}
$$

$\qquad \square$

### APPENDIX C
### POSITION IDENTIFICATION

We will analyze the singleton identification in two separate cases:

- $\mathscr{E}_{21}$: Event where the position is identified incorrectly when the bin is classified correctly a singleton
- $\mathscr{E}_{22}$: In the case of approximate matching, event where the position is identified incorrectly when the bin is originally a double-ton and one of the non-zero variable nodes has already been peeled off

**Definition 13** (Mutual Incoherence). *The mutual incoherence $\mu_{max}(\mathbf{W})$ of a matrix $\mathbf{W} = [\underline{w}_1 \ \underline{w}_2 \ \cdots \underline{w}_i \cdots \underline{w}_N]$ is defined as*

$$\mu_{max}(\mathbf{W}) \triangleq \max_{\forall i \neq j} \frac{|\underline{w}_i^\dagger \underline{w}_j|}{||\underline{w}_i|| \cdot ||\underline{w}_j||}$$

**Lemma 14** (Mutual Incoherence Bound for sub-sampled IDFT matrix [[6],Proposition A.1]). *The mutual incoherence $\mu_{max}(\mathbf{W_{i,k}})$ of the sensing matrix $\mathbf{W}_{i,k}$ (defined in Eq. 6), with B shifts, is upper bounded by*

$$\mu_{max} < 2\sqrt{\frac{\log(5N)}{B}}$$

*Proof.* The proof follows similar lines as the proof for Lemma V.3. in [6]. □

**Lemma 15.** *For the choice of $B = 4c_1^2 \log 5N$, the probability of error in identifying the position of a singleton at any bin $(i,j)$ can be upper bounded by*

$$\mathbb{P}[\mathscr{E}_{21}] \leq \exp\left\{ -\frac{N^{\mu-\alpha}(1-2\eta)^2(c_1^2-1)}{8(c_1^2+1)} \right\}$$

*Proof.* Let $j_p$ be the variable node participating in the singleton $(i,j)$. Then the observation vector $\underline{z}_{i,j}$ is given by

$$\underline{z}_{i,j} = \left[\underline{w}_{j_1}, \underline{w}_{j_2}, \quad \cdots \quad \underline{w}_{j_p}, \quad \cdots \quad \underline{w}_{j_{f_i}}\right] \times \begin{bmatrix} n_1 \\ \vdots \\ r[j_p] \\ \vdots \\ n_j \\ \vdots \\ n_{f_i} \end{bmatrix}$$

$$= r[j_p]\ \underline{w}_{j_p} + \sum_{k \neq p} n_k \underline{w}_{j_k}$$

where for convenience we use a simpler notation $j_k = j + (k-1)\frac{N}{f_i}, \underline{w}_{j_k} = \underline{w}^{j_k}$ as defined in Eq. and $n_l = \sum_{k=0}^{M-1} x[\theta_\ell + k]y[k]$ as defined in Eq. (10).

The estimated position $\hat{p}$ is given by

$$\hat{p} = \arg\max_l \ \frac{\underline{w}_{j_l}^\dagger \underline{z}_{i,j}}{B} \tag{11}$$

where † denotes the conjugate transpose of the vector. Also note that $||\underline{w}_{j_k}|| = B$ for any $j$ and $k$. From Eq. (11) we observe that the position is wrongly identified when $\exists p'$ such that

$$r[j_p] + \frac{1}{B}\sum_{k \neq p} n_k \underline{w}_{j_p}^\dagger \underline{w}_{j_k} \leq \frac{r[j_p]}{B}\ \underline{w}_{j_{p'}}^\dagger \underline{w}_{j_p} + n_{p'} + \frac{1}{B}\sum_{k \neq p, p'} n_k \underline{w}_{j_{p'}}^\dagger \underline{w}_{j_k}$$

$$\leftrightarrow \sum_{k \neq p, p'} \alpha_k n_k + \beta n_{p'} \geq r[j_p]\left(1 - \frac{\underline{w}_{j_{p'}}^\dagger \underline{w}_{j_p}}{B}\right) \geq M(1-2\eta)(1-\mu_{\max})$$

where $\alpha_k$ and $\beta$ are constants and can be shown to be in the range $\alpha_k \in [-2\mu_{\max}, 2\mu_{\max}]$ and $\beta \in [1 - \mu_{\max}, 1 + \mu_{\max}]$. Now using the bound given Chernoff Lemma in Lem. 6 we obtain

$$
\begin{aligned}
\mathbb{P}[\mathscr{E}_{21}] &\leq \exp\left\{ -\frac{2M(1-2\eta)^2(1-\mu_{\max})^2}{16 f_i \mu_{\max}^2 + 4(1+\mu_{\max})^2} \right\} \\
&\leq \exp\left\{ -\frac{2M(1-2\eta)^2(1-\mu_{\max})^2}{16(f_i \mu_{\max}^2 + 1)} \right\} \\
&\leq \exp\left\{ -\frac{2M(1-2\eta)^2(c_1-1)^2}{16(f_i + c_1^2)} \right\} \\
&\approx \exp\left\{ -\frac{N^{\mu-\alpha}(1-2\eta)^2(c_1^2-1)}{8(c_1^2+1)} \right\}
\end{aligned}
$$

The second inequality follows by the definition of $\mu_{\max} \leq 1$. We choose $B = 4c_1^2 \log 5N$, and substituting $\mu_{\max} \leq 2\sqrt{\frac{\log 5N}{B}} = 1/c_1$ (Lemma 14) we get the third inequality.

$\square$

**Lemma 16.** *For the choice of $B = 4c_1^2 \log 5N$, the probability of error in identifying the position of second non-zero variable node at a double-ton at any bin $(i,j)$, given that the first position identification is correct, can be upper bounded by*

$$
\mathbb{P}[\mathscr{E}_{22}] \leq \exp\left\{ -\frac{N^{\mu-\alpha}\,(c_1(1-2\eta)-1)^2}{8(1+c_1^2)} \right\}
$$

*Proof.* $\mathscr{E}_{22}$:

Let $j_p$ and $j_{\tilde{p}}$ be the two variable nodes participating in the doubleton $(i,j)$. Then the observation vector $\underline{z}_{i,j}$ is given by

$$
\underline{z}_{i,j} = \left[ \underline{w}_{j_1}, \underline{w}_{j_2}, \quad \cdots \quad \underline{w}_{j_p}, \quad \cdots \quad \underline{w}_{j_{f_i}} \right] \times
\begin{bmatrix}
n_1 \\
\vdots \\
r[j_p] \\
\vdots \\
n_j \\
\vdots \\
r[j_{\tilde{p}}] \\
\vdots \\
n_{f_i}
\end{bmatrix}
$$

$$
= r[j_p]\,\underline{w}_{j_p} + r[j_{\tilde{p}}]\,\underline{w}_{j_{\tilde{p}}} + \sum_{k \neq p} n_k \underline{w}_{j_k}
$$

Let the contribution from $j_{\tilde{p}}$ be peeled off from the doubleton at some iteration, then we get

$$
\underline{z}_{i,j} = r[j_p]\,\underline{w}_{j_p} + \frac{e_1}{B}\,\underline{w}_{j_{\tilde{p}}} + \sum_{k \neq p} n_k \underline{w}_{j_k}
$$

where $e_1 \in [-\eta M, \eta M]$ is an extra error term induced due to peeling off.

Now the estimated second position $\hat{p}$ is calculated using Eq. (11). We can observe that the position is wrongly identified when $\exists p'$ such that

$$
\frac{\underline{w}_{j_p}^\dagger \underline{z}_{i,j}}{B} \leq \frac{\underline{w}_{j_{p'}}^\dagger \underline{z}_{i,j}}{B}
$$

$$\implies r[j_p] + \frac{1}{B}\sum_{k\neq p,\tilde{p}} n_k \underline{\mathbf{w}}_{j_p}^\dagger \underline{\mathbf{w}}_{j_k} + \frac{e_1}{B}\underline{\mathbf{w}}_{j_p}^\dagger \underline{\mathbf{w}}_{j_{\tilde{p}}} \leq \frac{r[j_p]}{B}\ \underline{\mathbf{w}}_{j_{p'}}^\dagger \underline{\mathbf{w}}_{j_p} + n_{p'} + \frac{1}{B}\sum_{k\neq p,p',\tilde{p}} n_k \underline{\mathbf{w}}_{j_{p'}}^\dagger \underline{\mathbf{w}}_{j_k} + \frac{e_1}{B}\underline{\mathbf{w}}_{j_{p'}}^\dagger \underline{\mathbf{w}}_{j_{\tilde{p}}}$$

$$\leftrightarrow \sum_{k\neq p,p',\tilde{p}} \alpha_k n_k + \beta n_{p'} \geq r[j_p]\left(1 - \frac{\underline{\mathbf{w}}_{j_{p'}}^\dagger \underline{\mathbf{w}}_{j_p}}{B}\right) - \frac{2\eta M}{B}\underline{\mathbf{w}}_{j_{p'}}^\dagger \underline{\mathbf{w}}_{j_{\tilde{p}}}$$

$$\geq M(1-2\eta)(1-\mu_{\max}) - 2\eta M\mu_{\max} = M(1-2\eta-\mu_{\max})$$

where $\alpha_k$ and $\beta$ are constants and can be shown to be in the range $\alpha_k \in [-2\mu_{\max}, 2\mu_{\max}]$ and $\beta \in [1-\mu_{\max}, 1+\mu_{\max}]$. Now using the bound given by Chernoff Lemma in Lem. 6 we obtain

$$\mathbb{P}[\mathcal{E}_{22}] \leq \exp\left\{-\frac{2M(1-2\eta-\mu_{\max})^2}{16f_i\mu_{\max}^2 + 4(1+\mu_{\max})^2}\right\}$$

$$\leq \exp\left\{-\frac{2M(1-2\eta-\mu_{\max})^2}{16(f_i\mu_{\max}^2+1)}\right\}$$

$$\leq \exp\left\{-\frac{M(c_1(1-2\eta)-1)^2}{8(f_i+c_1^2)}\right\}$$

$$\leq \exp\left\{-\frac{N^{\mu-\alpha}\ (c_1(1-2\eta)-1)^2}{8(1+c_1^2)}\right\}$$

where for the choice of $B = 4c_1^2\log 5N$, $\mu_{\max} \leq 2\sqrt{\frac{\log 5N}{B}} = 1/c_1$.

$\square$

## Lemma 17.

## Lemma 18.

### REFERENCES

[1] A. Andoni, H. Hassanieh, P. Indyk, and D. Katabi, "Shift finding in sub-linear time," in *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 457–465, Society for Industrial and Applied Mathematics, 2013.

[2] A. Amir, M. Lewenstein, and E. Porat, "Faster algorithms for string matching with k mismatches," *Journal of Algorithms*, vol. 50, no. 2, pp. 257–275, 2004.

[3] R. S. Boyer and J. S. Moore, "A fast string searching algorithm," *Communications of the ACM*, vol. 20, no. 10, pp. 762–772, 1977.

[4] W. I. Chang and T. G. Marr, "Approximate string matching and local similarity," in *Annual Symposium on Combinatorial Pattern Matching*, pp. 259–273, Springer, 1994.

[5] H. Hassanieh, F. Adib, D. Katabi, and P. Indyk, "Faster gps via the sparse fourier transform," in *mobicom*, pp. 353–364, ACM, 2012.

[6] S. Pawar and K. Ramchandran, "A robust R-FFAST framework for computing a k-sparse n-length DFT in O($k\log n$) sample complexity using sparse-graph codes," in *Proc. IEEE Int. Symp. Information Theory*, pp. 1852–1856, 2014.