PRINCETON UNIVERSITY F'02 COS 597D: A THEORIST'S TOOLKIT

Lecture 5: Discrete Fourier Transform and its Uses

Lecturer: *Sanjeev Arora* Scribe:*Loukas Georgiadis*

1 Introduction

Sanjeev admits that he used to find fourier transforms intimidating as a student. His fear vanished once he realized that it is a rewording of the following trivial idea.
If $u_1, u_2, \ldots, u_n$ is an orthonormal basis of $\Re^n$ then every vector v can be expressed as $\sum_i \alpha_i u_i$ where $\alpha_i =< v, u_i >$ and $\sum_i \alpha_i^2 = |v|_2^2$.

Whenever the word "fourier transform" is used one is usually thinking about vectors as functions. In the above example, one can think of a vector in $\Re^n$ as a function from $\{0, 1, \ldots, n-1\}$ to $\Re$. Furthermore, the basis $\{u_i\}$ is defined in some nice way. Then the α_i's are called *fourier coefficients* and the simple fact mentioned about their ℓ_2 norm is called *Parseval's Identity*.

The classic fourier transform, using the basis functions $\cos(2\pi nx)$ and $\sin(2\pi nx)$, is analogous except we are talking about differentiable (or at least continuous) functions from $[-1, 1]$ to $\Re$, and the definition of "inner product" uses integrals instead of sums (the vectors are in an infinite dimensional space).

EXAMPLE 1 (Fast fourier transform) The FFT studied in undergrad algorithms uses the following orthonormal set $u_j = \left(1 \quad \omega_N^j \quad \ldots \quad \omega_N^{(N-1)j}\right)^T$ for $j = 0, 1, \ldots, N-1$. Here $\omega_N = e^{-2\pi i/N}$ is the N^{th} root of 1.

Consider a function $f : [1, N] \mapsto \mathbf{R}$, i.e. a vector in $\mathbf{R}^N$. The Discrete Fourier Transform (DFT) of f, denoted by $\hat{f}$, is defined as

$$\hat{f}_k = \sum_{x=1}^{N} f(x)\ \omega_N^{(k-1)(x-1)},\ k \in [1, N] \tag{1}$$

The inverse transform reconstructs the initial vector f by

$$f(x) = \frac{1}{N} \sum_{k=1}^{N} \hat{f}_k\ \omega_N^{-(k-1)(x-1)}. \tag{2}$$

In other words f is written as a linear combination of the basis vectors $u_j = \left(1 \quad \omega_N^j \quad \ldots \quad \omega_N^{(N-1)j}\right)^T$, for j in $[1, N]$. Another way to express this is by the matrix-vector product $\hat{f} = Mf$, where

$$M = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega_N^{1\cdot 1} & \omega_N^{1\cdot 2} & \dots & \omega_N^{1\cdot(N-1)} \\ 1 & \omega_N^{2\cdot 1} & \omega_N^{2\cdot 2} & \dots & \omega_N^{2\cdot(N-1)} \\ & & & \dots & \\ 1 & \omega_N^{(N-1)\cdot 1} & \omega_N^{(N-1)\cdot 2} & \dots & \omega_N^{(N-1)\cdot(N-1)} \end{pmatrix} \tag{3}$$

is the DFT matrix.

Fast Fourier Transform (FFT): We can exploit the special structure of M and divide the problem of computing the matrix-vector product Mf into two subproblems of size $N/2$. This gives an $O(N \log N)$ algorithm for computing the Fourier Transform. Using a simple property of the Discrete Fourier Transform and by applying the FFT algorithm to compute the DFT we can multiply n-bit numbers in $O(n \log n)$ operations.

2 Discrete Fourier Transform

In general we can use any orthonormal basis $u_1, u_2, \dots, u_N$. Then we can write

$$f = \sum_{i=1}^{N} \hat{f}_i u_i, \tag{4}$$

and the coefficients $\hat{f}_i$ are due to orthonormality equal to the inner product of f and of the corresponding basis vector, that is $\hat{f}_i =< f, u_i >$.

Definition 1 *The length of f is the L_2-norm*

$$\|f\|_2 = \sqrt{\sum_i f^2(i)}. \tag{5}$$

Parseval's Identity: The transform preserves the length of f, i.e.

$$\|f\|_2^2 = \sum_i \hat{f}_i^2. \tag{6}$$

2.1 Functions on the boolean hypercube

Now we consider real-valued functions whose domain is the boolean hypercube, i.e., $f : \{-1, 1\}^n \mapsto \mathbf{R}$. The inner product of two functions f and g is defined to be

$$< f, g >= \frac{1}{2^n} \sum_{\mathbf{x} \in \{-1,1\}^n} f(\mathbf{x}) g(\mathbf{x}). \tag{7}$$

Now we describe a particular orthonormal basis. For all the subsets S of $\{1,\ldots,n\}$ we define the functions

$$\chi_S(\mathbf{x}) = \prod_{i\in S} x_i. \tag{8}$$

where $\mathbf{x} = \begin{pmatrix} x_1 & \ldots & x_n \end{pmatrix}^T$ is a vector in $\{-1,1\}^n$. The χ_S's form an orthonormal basis. Moreover, they are the eigenvectors of the adjacency matrix of the hypercube.

EXAMPLE 2 Note that

$$\chi_\emptyset(\mathbf{x}) = 1$$

and

$$\chi_{\{i\}}(\mathbf{x}) = x_i.$$

Therefore,

$$< \chi_{\{i\}}, \chi_\emptyset >= \frac{1}{2^n} \sum_{\mathbf{x}\in\{-1,1\}^n} x_i = 0.$$

REMARK 1 Let $x_i = 1$ for all i in $I \subseteq \{1,\ldots,n\}$. Then $\chi_S(\mathbf{x}) = 1$ iff $|\overline{I} \cap S|$ is even.

REMARK 2 Now we show our basis is an orthonormal basis. Consider two subsets S and S' of $\{1,\ldots,n\}$. Then,

$$\begin{aligned} < \chi_S, \chi_{S'} > &= \frac{1}{2^n} \sum_{\mathbf{x}\in\{-1,1\}^n} \chi_S(\mathbf{x})\chi_{S'}(\mathbf{x}) = \frac{1}{2^n} \sum_{\mathbf{x}\in\{-1,1\}^n} \prod_{i\in S} x_i \prod_{j\in S'} x_j \\ &= \frac{1}{2^n} \sum_{\mathbf{x}\in\{-1,1\}^n} \prod_{i\in S\triangle S'} x_i =< \chi_{S\triangle S'}, \chi_\emptyset >= 0 \text{ unless } S = S'. \end{aligned}$$

Here $S \triangle S' = (S - S') \cup (S' - S)$ is the symmetric difference of S and S'.

Since the χ_S's form a basis every f can be written as

$$f = \sum_{S\subseteq\{1,\ldots,n\}} \hat{f}_S \chi_S, \tag{9}$$

where the coefficients are

$$\hat{f}_S =< f, \chi_S >= \frac{1}{2^n}\Big(\sum_{\mathbf{x}:\chi_S(\mathbf{x})=1} f(\mathbf{x}) - \sum_{\mathbf{x}:\chi_S(\mathbf{x})=-1} f(\mathbf{x})\Big). \tag{10}$$

REMARK 3 The basis functions are just the eigenvectors of the adjacency matrix of the boolean hypercube, which you calculated in an earlier homework. This observation —and its extension to Cayley graphs of groups— forms the basis of analysis of random walks on Cayley graphs using fourier transforms.

3 Applications of the Fourier Transform

In this section we describe two applications of the version of the Discrete Fourier Transform that we presented in section 2.1. The Fourier Transform is useful for understanding the properties of functions. The applications that we discuss here are: (i) PCP's and (ii) constructing small-bias probability spaces with a low number of random bits.

3.1 DFT and PCP's

We consider assignments of a boolean formula. These can be encoded in a way that enables them to be probabilistically checked by examining only a constant number of bits. If the encoded assignment satisfies the formula then the verifier accepts with probability 1. If no satisfying assignment exists then the verifier rejects with high probability.

DEFINITION 2 *The function $f : GF(2)^n \mapsto GF(2)$ is linear if there exist $a_1, \ldots, a_n$ such that $f(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i$ for all $\mathbf{x} \in GF(2)^n$.*

DEFINITION 3 *The function $g : GF(2)^n \mapsto GF(2)$ is δ-close if there exists a linear function f such that $\mathbf{Pr}_{x \in GF(2)^n}[f(\mathbf{x}) = g(\mathbf{x})] \geq 1 - \delta$.*

We think of g as an adversarily constructed function. The adversary attempts to deceive the verifier to accept g.

The functions that we considered are defined in $GF(2)$ but we can change the domain to be an appropriate group for the DFT of section 2.1.

$GF(2)$	$\{-1, 1\} \subseteq R$
$0 + 0 = 0$	$1 \cdot 1 = 1$
$0 + 1 = 1$	$1 \cdot (-1) = -1$
$1 + 1 = 0$	$(-1) \cdot (-1) = 1$

Therefore if we use the mapping $0 \mapsto 1$ and $1 \mapsto -1$ we have that $\sum_{i \in S} x_i \equiv \prod_{i \in S} x_i$. This implies that the linear functions become the Fourier Functions χ_S. The verifier uses the following test in order to determine if g is linear.

LINEAR TEST
Pick $\mathbf{x}, \mathbf{y} \in_R GF(2)^n$
Accept iff $g(\mathbf{x}) + g(\mathbf{y}) = g(\mathbf{x} + \mathbf{y})$

It is clear that a linear function is indeed accepted with probability one. So we only need to consider what happens if g is accepted with probability $1 - \delta$.

THEOREM 1
If g is accepted with probability $1-\delta$ then g is δ-close.

PROOF: We change our view from $GF(2)$ to $\{-1,1\}$. Notice now that

$$''\mathbf{x}+\mathbf{y}'' = \begin{pmatrix} x_1y_1 & x_2y_2 & \dots & x_ny_n \end{pmatrix}$$

and that if ρ is the fraction of the points in $\{-1,1\}^n$ where g and χ_S agree then by definition

$$\hat{g}_S =< g, \chi_S >= \rho - (1-\rho) = 2\rho - 1.$$

Now we need to show that if the test accepts g with probability $1-\delta$ it implies that there exists a subset S such that $\hat{g}_S \geq 1-2\delta$. The test accepts when for the given choice of $\mathbf{x}$ and $\mathbf{y}$ we have over $GF(2)$ that $g(\mathbf{x})+g(\mathbf{y}) = g(\mathbf{x}+\mathbf{y})$. So the equivalent test in $\{-1,1\}$ is $g(\mathbf{x})\cdot g(\mathbf{y})\cdot g(\mathbf{x}+\mathbf{y}) = 1$ because both sides must have the same sign.

Therefore if the test accepts with probability $1-\delta$ then

$$\mu = \mathbf{E}_{\mathbf{x},\mathbf{y}}[g(\mathbf{x})\cdot g(\mathbf{y})\cdot g(\mathbf{x}+\mathbf{y})] = (1-\delta)-\delta = 1-2\delta. \tag{11}$$

We replace g in (11) with its Fourier Transform and get

$$\begin{aligned}
\mu &= \mathbf{E}_{\mathbf{x},\mathbf{y}}\Big[\Big(\sum_{S_1\subseteq\{1,\dots,n\}} \hat{g}_{S_1}\chi_{S_1}(\mathbf{x})\Big)\cdot\Big(\sum_{S_2\subseteq\{1,\dots,n\}} \hat{g}_{S_2}\chi_{S_2}(\mathbf{y})\Big)\cdot\Big(\sum_{S_3\subseteq\{1,\dots,n\}} \hat{g}_{S_3}\chi_{S_3}(\mathbf{x}+\mathbf{y})\Big)\Big] \\
&= \mathbf{E}_{\mathbf{x},\mathbf{y}}\Big[\sum_{S_1,S_2,S_3\subseteq\{1,\dots,n\}} \hat{g}_{S_1}\ \hat{g}_{S_2}\ \hat{g}_{S_3}\ \ \chi_{S_1}(\mathbf{x})\ \chi_{S_2}(\mathbf{y})\ \chi_{S_3}(\mathbf{x}+\mathbf{y})\Big] \\
&= \sum_{S_1,S_2,S_3\subseteq\{1,\dots,n\}} \hat{g}_{S_1}\ \hat{g}_{S_2}\ \hat{g}_{S_3}\mathbf{E}_{\mathbf{x},\mathbf{y}}\Big[\chi_{S_1}(\mathbf{x})\ \chi_{S_2}(\mathbf{y})\ \chi_{S_3}(\mathbf{x}+\mathbf{y})\Big] \\
&= \sum_{S_1,S_2,S_3\subseteq\{1,\dots,n\}} \hat{g}_{S_1}\ \hat{g}_{S_2}\ \hat{g}_{S_3}\mathbf{E}_{\mathbf{x},\mathbf{y}}\Big[\chi_{S_1}(\mathbf{x})\ \chi_{S_3}(\mathbf{x})\ \chi_{S_2}(\mathbf{y})\ \chi_{S_3}(\mathbf{y})\Big].
\end{aligned}$$

However

$$\mathbf{E}_{\mathbf{x}}\Big[\chi_{S_1}(\mathbf{x})\chi_{S_3}(\mathbf{x})\Big] =< \chi_{S_1}, \chi_{S_3} >$$

which is non-zero only if $S_1 = S_3$. Similarly for fixed $\mathbf{x}$ we conclude that it must be $S_2 = S_3$. Therefore a non-zero expectation implies $S_1 = S_2 = S_3$ and we have

$$\mu = \sum_{S\subseteq\{1,\dots,n\}} \hat{g}_S^3 \leq \max_S|\hat{g}_S| \sum_{S\subseteq\{1,\dots,n\}} \hat{g}_S^2 = \max_S|\hat{g}_S|.$$

because Parseval's identity gives $\sum_S \hat{g}_S^2 = 1$. We conclude that $\max_S|\hat{g}_S| \geq 1-2\delta$.
□

3.2 Long Code

DEFINITION 4 *The function $f : GF(2)^n \mapsto GF(2)$ is a coordinate function if there is an $i \in \{1, \ldots, n\}$ such that $f(\mathbf{x}) = x_i$.*

Obviously a coordinate function is also linear.

DEFINITION 5 *The Long Code for $[1 \ldots W]$ encodes $w \in [1, \ldots, W]$ by the function $L_w(\mathbf{x}) = \mathbf{x}_w$.*

REMARK 4 A $\log W$ bit message is being encoded by a function from $GF(2)^W$ to $GF(2)$, i.e. with 2^W bits. So the Long Code is extremely redundant but is very useful in the context of PCP's.

Now the verifier is given a function and has to check if it is "close" to a long code. We give Hastad's test.

LONG CODE TEST
Pick $\mathbf{x}, \mathbf{y} \in_R GF(2)^W$
Pick a "noise" vector $\mathbf{z}$ such that $\mathbf{Pr}[z_i = 1] = \epsilon$
Accept iff $g(\mathbf{x}) + g(\mathbf{y}) = g(\mathbf{x} + \mathbf{y} + \mathbf{z})$

Suppose that $g(\mathbf{x}) = x_W$ for all $\mathbf{x}$. Then $g(\mathbf{x}) + g(\mathbf{y}) = x_W + y_W$, while $g(\mathbf{x} + \mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y} + \mathbf{z})_W = x_W + y_W + x_W$. So the probability of acceptance in this case is $1 - \epsilon$.

THEOREM 2 (HÅSTAD)
If the Long Word Test accepts with probability $1/2 + \delta$ then

$$\sum_S \hat{f}_S^3 (1 - 2\epsilon)^{|S|} \geq 2\delta.$$

The proof is similar to that of the Theorem 1 and is left as an exercise. The interpretation of Theorem 2 is that if f passes the test with must depend on a "few" coordinates. This is not the same as being close to a coordinate function but it suffices for the PCP application.

REMARK 5 Hastad's analysis belongs to a field called *influence of boolean variables*, that has been used recently in demonstration of sharp thresholds for many random graph properties (Friedgut and others) and in learning theory.

3.3 Reducing the random bits in small-bias probability spaces

We consider n random variables $x_1, \ldots, x_n$, where each x_i is in $\{0, 1\}$ and their joint probability distribution is D. Let U be the uniform distribution. The distance of D from the uniform distribution is

$$\|D - U\| = \sum_{\omega \in \{0,1\}^n} |U(\omega) - D(\omega)|.$$

DEFINITION 6 *The bias of a subset $S \subseteq \{1, \ldots, n\}$ for a distribution D is*

$$bias_D(S) = \Big|\mathbf{Pr}_D\Big[\sum_{i\in S} x_i = 1\Big] - \mathbf{Pr}_D\Big[\sum_{i\in S} x_i = 0\Big]\Big|. \tag{12}$$

Here we are concerned with the construction of (small) ϵ-bias probability spaces that are k-wise independent, where k is a function of ϵ. Such spaces have several applications. For example they can be used to reduce the amount of randomness that is necessary in several randomized algorithms and for derandomizations.

DEFINITION 7 *The variables $x_1, \ldots, x_n$ are ϵ-biased if for all subsets $S \subseteq \{1, \ldots, n\}$, $bias_D(S) \leq \epsilon$. They are k-wise ϵ-biased if for all subsets $S \subseteq \{1, \ldots, n\}$ such that $|S| \leq k$, $bias_D(S) \leq \epsilon$.*

DEFINITION 8 *The variables $x_1, \ldots, x_n$ are k-wise δ-dependent if for all subsets $S \subseteq \{1, \ldots, n\}$ such that $|S| \leq k$, $\|U(S) - D(S)\| \leq \delta$. ($D(S)$ and $U(S)$ are the distributions D and U restricted to S.)*

REMARK 6 If $\delta = 0$ this is just the notion of k-wise independence studied briefly in an earlier lecture.

The following two conditions are equivalent:

1. All x_i are independent and $\mathbf{Pr}[x_i = 0] = \mathbf{Pr}[x_i = 1]$.
2. For every subset $S \subseteq \{1, \ldots, n\}$ we have $\mathbf{Pr}[\sum_{i\in S} x_i = 0] = \mathbf{Pr}[\sum_{i\in S} x_i = 1]$, that is the parity of the subset is equally likely to be 0 or 1.

Here we try to reduce the size of the sample space by relaxing the second condition, that is we consider biased spaces.

We will use the following theorem.

THEOREM 3 (DIACONIS AND SHAHSHAHANI '88)

$$\|D - U\| \leq 2^n \Big(\sum_{S\subseteq\{1,\ldots,n\}} \hat{D}_S^2\Big)^{1/2} = \Big(\sum_{S\subseteq\{1,\ldots,n\}} bias_D(S)^2\Big)^{1/2}. \tag{13}$$

PROOF: We write the Fourier expansion of D as $D = \sum_S \hat{D}_S \chi_S$. Then,

$$D = U + \sum_{S\neq\emptyset} \hat{D}_S \chi_S. \tag{14}$$

Now observe that

$$\|\sum_{S\neq\emptyset} \hat{D}_S \chi_S\|_2^2 = 2^n \Big\langle \sum_{S\neq\emptyset} \hat{D}_S \chi_S, \sum_{S\neq\emptyset} \hat{D}_S \chi_S \Big\rangle = 2^n \sum_{S\neq\emptyset} \hat{D}_S^2 \leq 2^n \sum_S \hat{D}_S^2. \tag{15}$$

Also by the Cauchy-Schwartz inequality

$$\|D - U\| = \|D - U\|_1 \leq 2^{n/2}\|D - U\|_2. \tag{16}$$

Thus

$$\|D - U\| \leq 2^n \Big(\sum_{S \subseteq \{1,\ldots,n\}} \hat{D}_S^2 \Big)^{1/2} = \Big(\sum_{S \subseteq \{1,\ldots,n\}} bias_D(S)^2 \Big)^{1/2}. \tag{17}$$

The second equality holds from equation (10). □

If we restrict our attention to subsets of size at most k then Theorem 13 implies:

Corollary 4
If D is ϵ-biased then it is also k-wise δ-independent for $\delta = 2^{k/2}\epsilon$.

The main theorem that we want to show is

Theorem 5 (J. Naor and M. Naor)
We can construct a probability space of ϵ-biased $\{0,1\}$ random variables $x_1, \ldots, x_n$ using $O(\log n + \log \frac{1}{\epsilon})$ random bits.

Proof: The construction consists of three stages:

Stage 1: Construct a family $\mathcal{F} \subseteq \{0,1\}^n$ such that for $\mathbf{r} \in_R \mathcal{F}$ and for all subsets $S \subseteq \{1, \ldots, n\}$,

$$\mathbf{Pr}[\chi'_S(\mathbf{r}) = 1] \geq \beta,$$

where

$$\chi'_S(\mathbf{x}) = \sum_{i \in S} x_i \ (\mathrm{mod}\ 2).$$

and β is some constant. We discuss this stage in section 3.3.1.

Stage 2: Sample $\ell = O(\log(1/\epsilon))$ vectors $\mathbf{r}^1, \ldots, \mathbf{r}^\ell$ from $\mathcal{F}$ such that for all subsets S

$$\mathbf{Pr}_{\mathbf{r}^1,\ldots,\mathbf{r}^\ell}\Big[\forall i \in \{1,\ldots,\ell\}, \chi'_S(\mathbf{r}^i) = 0\Big] \leq \epsilon.$$

Note that if we just sample uniformly from $\mathcal{F}$ with $\log(1/\epsilon)$ samples then we need $\log n \cdot \log(1/\epsilon)$ random bits. We can do better by using a random walk on an expander. This will reduce the number of necessary random bits to $O(\log n + \log(1/\epsilon))$.

Stage 3: Let $\mathbf{a}$ be a vector in $\{0,1\}^\ell$ chosen uniformly at random. The assignment to the random variables $x_1, \ldots, x_n$ is

$$\begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix}^T = \sum_{i=1}^{\ell} a_i \mathbf{r}^i$$

With probability at least $1 - \epsilon$ not all $\mathbf{r}^i$'s are such that for all S, $\chi'_S(\mathbf{r}^i) = 0$. But if there exists a vector $\mathbf{r}^j$ in the sample with $\chi'_S(\mathbf{r}^j) = 1$ then we can argue that $\mathbf{Pr}_D[\chi'_S(x_1, \ldots, x_n) = 1] = \mathbf{Pr}_D[\chi'_S(x_1, \ldots, x_n) = 0]$. Consequently $bias_D(S) \leq \epsilon$. □

3.3.1 Construction of the family $\mathcal{F}$

We construct $\log n$ families $\mathcal{F}_1, \ldots, \mathcal{F}_{\log n}$. For a vector $\mathbf{v} \in \{0,1\}^n$ let $|\mathbf{v}|$ be the number of ones in $\mathbf{v}$. For a vector $\mathbf{r}$ chosen uniformly at random from $\mathcal{F}_i$ and for each vector $\mathbf{v}$ in $\{0,1\}^n$ with $k \leq |\mathbf{v}| \leq 2k-1$ we require that

$$\mathbf{Pr}_{\mathbf{r}}[< \mathbf{r}, \mathbf{v} >= 1] \geq \beta.$$

Assume that we have random variables that are uniform on $\{1, \ldots, n\}$ and c-wise independent (using only $O(c \log n)$ random bits - see Lecture 4). We define three random vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{r}$ such that

1. The u_i's are c-wise independent and $\mathbf{Pr}[u_i = 0] = \mathbf{Pr}[u_i = 1] = 1/2$.
2. The w_i's are pairwise independent and $\mathbf{Pr}[w_i = 1] = 2/k$.
3. The elements of $\mathbf{r}$ are

$$r_i = \begin{cases} 1 & \text{if } u_i = 1 \text{ and } w_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

We claim that $\mathbf{Pr}[< \mathbf{r}, \mathbf{v} >= 1] \geq 1/4$ for all $\mathbf{v}$ with $k \leq |\mathbf{v}| \leq 2k-1$. In order to prove this we define the vector $\mathbf{v}'$ with coordinates:

$$v'_i = \begin{cases} 1 & \text{if } v_i = 1 \text{ and } w_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

By the above definition we have $< \mathbf{v}', \mathbf{u} >=< \mathbf{r}, \mathbf{v} >$. We wish to show that

$$\mathbf{Pr}[1 \leq |\mathbf{v}'| \leq c] \leq \frac{1}{2}. \tag{18}$$

Because the u_i's are c-wise independent (18) sufficies to guarantee that $\mathbf{Pr}[< \mathbf{v}', \mathbf{u} >= 1] \geq 1/4$. We view $|\mathbf{v}'|$ as a binomial random variable h with probability $p = 2/k$ for each non-zero entry. Therefore, $\mathbf{E}[h] = pl$ and $\mathbf{Var}[h] = p(1-p)l$, where $l \in [k, 2k-1]$ is the number of potential non-zero entries. By Chebyshev's inequality at the endpoints of $[k, 2k-1]$ we have

$$\mathbf{Pr}[|h - pk| \geq 2] \leq \frac{pk(1-p)}{4} \leq \frac{1}{2}$$

and

$$\mathbf{Pr}[|h - 2pk| \geq 3] \leq \frac{2pk(1-p)}{9} \leq \frac{4}{9}.$$

Thus, for $c = 2kp + 3 = 7$ we get success probability at least $1/2$.

So if $|\mathbf{v}|$ is approximately known we can get high probability of success with $O(\log n)$ bits. Since we don't know $|\mathbf{v}|$ we chose $a \in_R \{0,1\}^{\log n}$ and construct

$$\mathbf{r} = \sum_{i=1}^{\log n} a_i \mathbf{r}^i, \text{ where } \mathbf{r}^i \in_R \mathcal{F}_i.$$

Each family $\mathcal{F}_i$ corresponds to the case that $|\mathbf{v}| \in [2^{i-1}, 2^i - 1]$. Now for any $\mathbf{v}$ we can show that $\mathbf{Pr}[< \mathbf{v}, \mathbf{r} >= 1] \geq \frac{1}{8}$ (so $\beta = 1/8$).