Bucketing Coding and Information Theory for the Statistical High Dimensional Nearest Neighbor Problem

Moshe Dubiner

***Abstract*—The problem of finding high dimensional approximate nearest neighbors is considered, when the data is generated by some known probabilistic model. A large natural class of algorithms (bucketing codes) is investigated, Bucketing information is defined, and is proven to bound the performance of all bucketing codes. The bucketing information bound is asymptotically attained by some randomly constructed bucketing codes.**

The example of n Bernoulli(1/2) very long (length $d \to \infty$) sequences of bits is singled out. It is assumed that $n-2m$ sequences are completely independent, while the remaining $2m$ sequences are composed of m dependent pairs. The interdependence within each pair is that their bits agree with probability $1/2 < p \leq 1$. It is well known how to find most pairs with high probability by performing order of $n^{\log_2 2/p}$ comparisons. It is shown that order of $n^{1/p+\epsilon}$ comparisons suffice, for any $\epsilon > 0$. A specific two dimensional inequality (proven in another paper) implies that the exponent $1/p$ cannot be lowered. Moreover if one sequence out of each pair belongs to a known set of $n^{(2p-1)^2}$ sequences, pairing can be done using order $n^{1+\epsilon}$ comparisons!

***Index Terms*—Approximate nearest neighbor, Information theory**

I. Introduction

Suppose we have two bags of points, X_0 and X_1, randomly distributed in a high-dimensional space. The points are independent of each other, with the exception that some unknown matched pairs from $X_0 \times X_1$ are significantly "closer" to each other than random chance would account for. We want an efficient algorithm for quickly finding these pairs. We worked on finding texts that are translations of each other, which is a two bags problem (the bags are languages). In most cases there is only one bag $X_0 = X_1 = X$, $n_0 = n_1 = n$. The two bags model is slightly more complicated, but leads to clearer thinking. It is a bit reminiscent of fast matrix multiplication: even when one is interested only in square matrices, it pays to consider rectangular matrices too.

Let us start with the well known simple uniform marginally Bernoulli(1/2) example. Suppose $X_0, X_1 \subset \{0,1\}^d$ of sizes n_0, n_1 respectively are randomly chosen as independent Bernoulli(1/2) variables, with one exception. Choose uniformly randomly one point $x_0 \in X_0$, xor it with a random Bernoulli$(1-p)$ vector and overwrite one uniformly chosen random point $x_1 \in X_1$. A symmetric description is to say that x_0, x_1 i'th bits have the joint probability matrix

$$P_i = \begin{pmatrix} p/2 & (1-p)/2 \\ (1-p)/2 & p/2 \end{pmatrix} \tag{1}$$

for some known $1/2 < p \leq 1$. In practice p will have to be estimated. What does information theory say? For simplicity let us consider a single matched pair. It is distinguished by having a smaller than expected Hamming distance. Define

$$\ln W = \ln n_0 + \ln n_1 - \sum_{i=0}^{d} I(P_i) \tag{2}$$

where $I(P_i)$ is the mutual information between the matched pair's i'th coordinate values ($p\ln(2p)+(1-p)\ln(2(1-p))$ for the example). Information theory tells us that we can not hope to pin the matched pair down into less than W possibilities, but can come close to it in some asymptotic sense. Assume that W is small. How can we find the matched pairs? The trivial way to do it is to compare all the $n_0 n_1$ pairs. Many papers have shown how to do this in not much more than $O(n_0+n_1)$ operations, when the dimension d is fixed. However W small implies that d is at least of order $\ln(n_0+n_1)$. There is some literature dealing with the high dimensional nearest neighbor problem. The earliest references I am aware of are Manber [10], Paturi, Rajasekaran and Reif [12], Karp,Waarts and Zweig [9], Broder [3], Indyk and Motwani [8]. They do not limit themselves to our simplistic example, but can handle it. Without restricting generality let $n_0 \leq n_1$. Randomly choose

$$m \approx \log_2 n_0 \tag{3}$$

out of the d coordinates, and compare the point pairs which agree on these coordinates (in other words, fall into the same bucket). The expected number of comparisons is at most

$$n_0 n_1 2^{-m} \approx n_1 \tag{4}$$

while the probability of success of one comparison is p^m. In case of failure try again, with other random m coordinates. At first glance it might seem that the expected number of tries until success is p^{-m}, but that is not true because the attempts are interdependent. An extreme example is $d = m$, where the attempts are identical. In the unlimited data case $d \to \infty$ the expected number of tries is indeed p^{-m}, so the expected number of comparisons is

$$W \approx p^{-m} n_1 \approx n_0^{\log_2 1/p} n_1 \tag{5}$$

Is this optimal? It seems that when the dimension d is large one should be able to profit by cherry picking some "good" parts of

M. Dubiner is with Google, e-mail: moshe@google.com

Manuscript submitted to IEEE Transactions on Information Theory on March 3 ,2007; revised on February 20, 2009 and January 20, 2010.

the data. For instance, divide the coordinates into triplets and replace each triplet by its majority. This particular scheme does not work (it effectively replaces p with the lower $p^3+3p(1-p)^2$). Other simple attempts fail too. Alon [1] has suggested the possibility of improvement for $p = 1 - o(1)$ by using Hamming's perfect code.

In this article, we prove that in the $n_0 = n_1 = n$ case, $W \approx n^{\log_2 2/p}$ can be reduced to

$$W \approx n^{1/p+\epsilon} \tag{6}$$

for any $1/2 < p < 1$, $\epsilon > 0$. The algorithm is described in the next section. Amazingly it is possible to characterize the asymptotically best exponent not only for this problem, but for a much larger class. We allow non binary discrete data, a limited amount of data ($d < \infty$) and a general probability distribution of each coordinate. It turn out that

$$\ln W \geq \sup_{\lambda_0,\lambda_1 \leq 1 \leq \mu, \lambda_0+\lambda_1} \left[\lambda_0 \ln n_0 + \lambda_1 \ln n_1 + \mu \ln S - \right. \\ \left. - \sum_{i=1}^{d} I(P_i, \lambda_0, \lambda_1, \mu) \right] \tag{7}$$

where W is the work, S is the success probability and $I(P_i, \lambda_0, \lambda_1, \mu)$ is a newly defined **bucketing information** function, generalizing Shanon's mutual information $I(P) = I(P_i, 1, 1, \infty)$. We prove that the inequality is asymptotically tight, see section V. Full general proofs are given in the appendices. The rest of this paper contains instructive less general proofs and examples.

II. An Asymptotically Better Algorithm

The following algorithm works for the uniform marginally Bernoulli(1/2) problem (1) with $1/2 < p < 1$. Let $0 < d_0 \leq d$ be some natural numbers. We construct a d dimensional bucket in the following way. Choose a random point $b \in \{0,1\}^d$. The bucket contains all points $x \in \{0,1\}^d$ such for exactly $d_0 - 1$ or d_0 coordinates i, we have $x_i = b_i$. (It is even better to allow $d_0 - 1, \ldots, d$, but the analysis gets a little messy.) The algorithm uses T such buckets, independently chosen. All point pairs falling into the same bucket are compared. A true matching pair is considered successfully found iff it gets compared. The probability of a point x falling into a bucket is

$$p_{A*} = \binom{d}{d_0 - 1} 2^{-d} + \binom{d}{d_0} 2^{-d} \tag{8}$$

Let the number of points be

$$n_0 = n_1 = n = \lfloor 1/p_{A*} \rfloor \tag{9}$$

This way the expected number of comparisons (point pairs in the same bucket) is at most

$$T(np_{A*})^2 \leq T \tag{10}$$

The probability that a matched pair falls at least once into the same bucket is

$$S = \sum_{m=0}^{d} \binom{d}{m} p^{d-m}(1-p)^m \left[1 - (1 - S_m)^T\right] \tag{11}$$

$$S_m = 2^{-d} \binom{m}{\lfloor m/2 \rfloor} \cdot \\ \cdot \left[\binom{d-m}{d_0 - \lceil m/2 \rceil} + \binom{d-m}{d_0 - \lceil (m+1)/2 \rceil} \right] \tag{12}$$

The explanation follows. In these formulas m is the number of coordinates i at which the matched pair's values disagree: $x_{0,i} \neq x_{1,i}$. Consider the matched pair fixed. There are 2^d possible buckets, independently chosen. Consider one bucket. For $j, k = 0, 1$ denote by m_{jk} the number of coordinates i such that $x_{0,i} \oplus b_i = j$ and $x_{0,i} \oplus x_{1,i} = k$ where $\oplus$ is the xor operation. We know that $m_{01} + m_{11} = m$ and $m_{00} + m_{10} = d - m$. Both x_0, x_1 fall into the basket iff $m_{00} + m_{01} = d_0 - 1, d_0$ and $m_{00} + m_{11} = d_0 - 1, d_0$. There are two possibilities

$$\begin{pmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{pmatrix} = \begin{pmatrix} d_0 - \lceil m/2 \rceil & \lfloor m/2 \rfloor \\ d - d_0 - \lfloor m/2 \rfloor & \lceil m/2 \rceil \end{pmatrix} \tag{13}$$

$$\begin{pmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{pmatrix} = \begin{pmatrix} d_0 - \lceil (m+1)/2 \rceil & \lceil m/2 \rceil \\ d - d_0 - \lfloor (m-1)/2 \rfloor & \lfloor m/2 \rfloor \end{pmatrix} \tag{14}$$

each providing $\binom{m_{00}+m_{10}}{m_{00}} \binom{m_{01}+m_{11}}{m_{01}}$ buckets.

Clearly m obeys a Binomial$(m, 1-p)$ distribution, so by Chebyshev's inequality and $(1 - S_m)^T \leq e^{-TS_m}$, for any $\delta > 0$

$$S \geq \min_{|m-(1-p)d| < \sqrt{p(1-p)d/\delta}} \left(1 - e^{-TS_m} - \delta\right) \tag{15}$$

Hence taking

$$T = \lceil -\ln \delta / \min_{|m-(1-p)d| < \sqrt{p(1-p)d/\delta}} S_m \rceil \tag{16}$$

guaranties a success probability $S \geq 1 - 2\delta$. What is the relationship between n and T? Let

$$d_0 \sim (1+\delta)d/2, \quad d \to \infty \tag{17}$$

By Stirling's approximation

$$\lim_{d\to\infty} \frac{\ln n}{d} = \ln 2 - H\left(\frac{1+\delta}{2}\right) = \\ = \frac{1+\delta}{2}\ln(1+\delta) + \frac{1-\delta}{2}\ln(1-\delta) \tag{18}$$

and

$$\lim_{d\to\infty} \frac{\ln T}{d} = p\ln 2 - pH\left(\frac{1+\delta/p}{2}\right) = \\ = \frac{p+\delta}{2}\ln(1+\delta/p) + \frac{p-\delta}{2}\ln(1-\delta/p) \tag{19}$$

Letting $\delta \to 0$ results in exponent

$$\lim_{\delta\to 0}\lim_{d\to\infty} \frac{\ln T}{\ln n} = \frac{1}{p} \tag{20}$$

We are not yet finished with this algorithm, because the number of comparisons is not the only component of work. One also has to throw the points into the buckets. The straightforward way of doing it is to check the point-bucket pairs. This involves $2nT$ checks, which is worse than the naive n^2 algorithm! This is overcome as follows. Suppose we want to handle $\tilde{n}$ points having $\tilde{d}$ coordinates. Let $n = \lceil \tilde{n}^{1/k} \rceil$ and $d = \lfloor \tilde{d}/k \rfloor$ for some $k \geq 1$, and take the k'th tensor power

of the previous algorithm. That means throwing n^k points in $\{0,1\}^{kd}$ into T^k buckets. The success probability is S^k, the expected number of comparisons is at most T^k, but throwing the points into the buckets takes only an expected number of $2n^kT$ vector operations (of length kd). Hence the total expected number of vector operations is at most

$$T^k + 2n^kT \tag{21}$$

The bucketing is done by depth first search over the regular T branching tree of depth k. Vertex $(t_1, t_2, \ldots, t_j)$ at depth j points to those original n^k points which fell into bucket $0 \le t_1 < T$ by their first d coordinates etc. Hence

$$k + kn^k \tag{22}$$

pointers suffice to represent the path from the root to the current vertex and its content. Taking

$$k = \lceil 1/(1-p) \rceil \tag{23}$$

lets us approach the promised (6) exponent $1/p$.

III. The Probabilistic Approximate Nearest Neighbor Problem

How should the approximate nearest neighbor problem be modeled? We follow in the footsteps of noisy channel theory.

Definition 3.1: Let the sets

$$X_0 \subset \{0,1,\ldots,b_0-1\}^d, \quad X_1 \subset \{0,1,\ldots,b_1-1\}^d \tag{24}$$

of cardinalities $\#X_0 = n_0$, $\#X_1 = n_1$ be randomly constructed using probability matrices

$$P_i = \begin{pmatrix} p_{i,00} & p_{i,01} & \cdots & p_{i,0\ b_1-1} \\ p_{i,10} & p_{i,11} & \cdots & p_{i,0\ b_1-1} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i,b_0-1\ 0} & p_{i,b_0-1\ 1} & \cdots & p_{i,b_0-1\ b_1-1} \end{pmatrix} \tag{25}$$

$$p_{i,jk} \ge 0, \quad p_{i,**} = \sum_{j=0}^{b_0-1} \sum_{k=0}^{b_1-1} p_{i,jk} = 1 \tag{26}$$

The $n_0 n_1$ vector pairs are divided into 3 (possibly empty) classes: matched, random and monochromatic. For a matched pair $x \in X_0,\ y \in X_1$ the probability that $x_i = j$ and $y_i = k$ is $p_{i,jk}$ and there is independence between coordinates. For any set $B \subset \{0,\ldots,b_0-1\}^d \times \{0,\ldots,b_1-1\}^d$

$$p_B = \sum_{(x,y)\in B} \prod_{i=1}^{d} p_{i,x_i y_i} \tag{27}$$

For a random pair $x \in X_0,\ y \in X_1$ the probability that $x_i = j$ and $y_i = k$ is $p_{i,j*}p_{i,*k}$ where the marginal probabilities are

$$p_{i,j*} = \sum_{k=0}^{b_1-1} p_{i,jk} \qquad p_{i,*k} = \sum_{j=0}^{b_0-1} p_{i,jk} \tag{28}$$

and there is independence between coordinates. For any $B_0 \subset X_0,\ B_1 \subset X_1$

$$p_{B_0 *} = \sum_{x\in B_0} \prod_{i=1}^{d} p_{x_i *} \qquad p_{*B_1} = \sum_{y\in B_1} \prod_{i=1}^{d} p_{*y_i} \tag{29}$$

Without restricting generality we require $p_{i,j*}p_{i,*k} > 0$ (otherwise remove the zero rows and columns). Monochromatic pairs are allowed only when $p_{i,j*} = p_{i,*j}$, for the technical purpose of duplicating a monochromatic set of points $X_0 = X_1 = X$. The probability that $x_i = y_i = j$ is $p_{i,j*}$, with independence between coordinates.

How much work is necessary in order to find most matched pairs? For general algorithms, this seems an extremely hard question. All known algorithms rely on collecting putative point pairs in buckets of some sort, so we will limit ourselves to these algorithms. But what are bucketing algorithms? One could choose a X_0 point and a X_1 point in some complicated data dependent way, then throw them into a single bucket. It is unlikely to work, but can you prove it? In order to disallow such knavery we will insist on data independent buckets. Most practical bucketing algorithms are data dependent. That is necessary because the data is used to construct (usually implicitly) a data model. We suspect that when the data model is known, there is little to be gained by making the buckets data dependent, but would welcome and greatly appreciate even a single contrary example.

Definition 3.2: A bucketing code is a set of T subset pairs

$$(B_{0,0}, B_{1,0}), \ldots, (B_{0,T-1}, B_{1,T-1}) \subset X_0 \times X_1 \tag{30}$$

Its success probability is

$$S = p_{\cup_{t=0}^{T-1} B_{0,t}\times B_{1,t}} \tag{31}$$

and for any real numbers $n_0, n_1 > 0$ its work is

$$W = \max\left(\sum_{t=0}^{T-1} n_0 p_{B_{0,t}*}\ ,\ \sum_{t=0}^{T-1} n_1 p_{*B_{1,t}}\ ,\ \sum_{t=0}^{T-1} n_0 p_{B_{0,t}*} n_1 p_{*B_{1,t}} \right) \tag{32}$$

Clearly S is the probability that a matched pair falls together into at least one bucket. Work has to be explained. In the above definition we consider n_0, n_1 to be the expected number of X_0, X_1 points, so they are not necessarily integers. The simplest implementation of a bucketing code is to store it as two arrays of lists, indexed by the possible vectors. The first array of size b_0^d keeps for each point $x \in \{0,1,\ldots,b_0-1\}^d$ the list of buckets (from 0 to $T-1$) which contain it. The second array of size b_1^d does the same for the $B_{1,t}$'s. When we are given X_0 and X_1 we look each element up, and accumulate pointers to it in linked lists, each originating from a bucket. Then we compare the pairs in each of the T buckets. Let us count the expected number of operations. The expected number of buckets containing any specific X_0 point is $\sum_{t=0}^{T-1} p_{B_{0,t}*}$, so the X_0 lookups involve an order of $n_0 + \sum_{t=0}^{T-1} p_{B_{0,t}*}$ operations. Similarly the X_1 lookups take $n_1 + \sum_{t=0}^{T-1} p_{*B_{1,t}}$ The probability that a specific random pair falls into bucket t is $p_{B_{0,t}*}p_{*B_{1,t}}$, so the expected number of comparisons is $n_0 p_{B_{0,t}*} n_1 p_{*B_{1,t}}$ It all adds up to at most $n_0 + n_1 + 3W$.

The behavior of of S and W under tensor product is important. Suppose we have two bucketing codes defined on different coordinates. Concatenating the codes gives success

probability of exactly S_1S_2 (success probability is multiplicative), while the work if bounded from above by W_1W_2 (work is sub-multiplicative). There is no value in having $W < \max(n_0, n_1)$ by itself, but it is advantageous to allow it in a tensor product factor. For that reason formula (32) does not include n_0 and n_1 inside the maximization brackets.

The fly in the ointment is that for even moderate dimension d the memory requirement is out of the universe. Hence the code in memory approach can be used only for small d. This is familiar from the theory of block coding. Higher dimensions can be handled by splitting them up into short blocks (as in the previous section), or by more sophisticated coding algorithms. In any case W is a lower bound on the actual work of a bucketing algorithm.

IV. Bucketing Information

We will show in the next section that the performance of bucketing codes is governed by the **bucketing information** function $I(P, \lambda_0, \lambda_1, \mu)$, defined for

$$0 \le \lambda_0, \lambda_1 \le 1 \le \mu, \lambda_0 + \lambda_1 \tag{33}$$

We have an embarrassment of riches: three different bucketing information formulas (not counting in-between variations), none of them intuitive. The shortest is still pretty long

$$I(P, \lambda_0, \lambda_1, \mu) = \min_{\substack{\{z_{jk}\ge 0\}_{jk} \\ 0\le j<b_0 \\ 0\le k<b_1 \\ z_{**}=1}} \max_{\substack{\{x_{0,j}\ge 0\}_j \\ 0\le j<b_0 \\ x_{0,*}=1}} \max_{\substack{\{x_{1,k}\ge 0\}_k \\ 0\le k<b_1 \\ x_{1,*}=1}} \ln \sum_{j=0}^{b_0-1}\sum_{k=0}^{b_1-1} p_{jk}\left(\frac{p_{jk}}{z_{jk}}\right)^{\mu-1}\left(\frac{x_{0,j}}{p_{j*}}\right)^{\lambda_0}\left(\frac{x_{1,k}}{p_{*k}}\right)^{\lambda_1} \tag{34}$$

Defining a $b_0 \times b_1$ matrix $M(P, Z, \lambda_0, \lambda_1, \mu)$ by

$$m_{jk} = p_{jk}\left(\frac{p_{jk}}{z_{jk}}\right)^{\mu-1}(p_{j*})^{-\lambda_0}(p_{*k})^{-\lambda_1} \tag{35}$$

and denoting by $\|\ \|_{\alpha,\beta}$ the operator norm from L_α to L_β, formula (34) can be written as

$$I(P, \lambda_0, \lambda_1, \mu) = \min_Z \ln \|M(P, Z, \lambda_0, \lambda_1, \mu)\|_{\frac{1}{\lambda_0}, \frac{1}{1-\lambda_1}} \tag{36}$$

where Z goes over all probability matrices. In order to see some connection with classical information theory let us consider $\lambda_0 = \lambda_1 = 1$. The x maximization is easy, shifting all weight to a single cell:

$$I(P, 1, 1, \mu) = \ln \min_{\substack{\{z_{jk}\ge 0\}_{jk} \\ 0\le j<b_0 \\ 0\le k<b_1 \\ z_{**}=1}} \max_{\substack{0\le j<b_0 \\ 0\le k<b_1}} \frac{p_{jk}}{p_{j*}p_{*k}}\left(\frac{p_{jk}}{z_{jk}}\right)^{\mu-1} \tag{37}$$

Hence $\frac{p_{jk}}{p_{j*}p_{*k}}\left(\frac{p_{jk}}{z_{jk}}\right)^{\mu-1} \le e^{I(P,1,1,\mu)}$ so $z_{jk} \ge p_{jk}\left(\frac{p_{jk}}{p_{j*}p_{*k}}e^{-I(P,1,1,\mu)}\right)^{\frac{1}{\mu-1}}$ which together with $z_{**} = 1$ gives $1 \ge \sum_{jk} p_{jk}\left(\frac{p_{jk}}{p_{j*}p_{*k}}\right)^{\frac{1}{\mu-1}} e^{-\frac{I(P,1,1,\mu)}{\mu-1}}$. Equality is clearly achievable so

$$I(P, 1, 1, \mu) = (\mu - 1)\ln \sum_{j=0}^{b_0-1}\sum_{k=0}^{b_1-1} p_{jk}\left(\frac{p_{jk}}{p_{j*}p_{*k}}\right)^{\frac{1}{\mu-1}} \tag{38}$$

In particular for $\mu \to \infty$ we have $\left(\frac{p_{jk}}{p_{j*}p_{*k}}\right)^{\frac{1}{\mu-1}} = 1 + \frac{1}{\mu-1}\ln\left(\frac{p_{jk}}{p_{j*}p_{*k}}\right) + O\left((\mu-1)^{-2}\right)$ so

$$I(P, 1, 1, \infty) = I(P) = \sum_{j=0}^{b_0-1}\sum_{k=0}^{b_1-1} p_{jk}\ln\frac{p_{jk}}{p_{j*}p_{*k}} \tag{39}$$

where $I(P)$ is Shannon's mutual information.

Formula (34) does not make it clear that $I(P, \lambda_0, \lambda_1, \mu)$ is nonnegative, far less its other basic properties. We will prove in appendix A that

Theorem 4.1: For any probability matrix P the bucketing information function $I(P, \lambda_0, \lambda_1, \mu)$ is nonnegative, monotonically nondecreasing and convex in $\lambda_0, \lambda_1, -\mu$. Moreover

$$I(P, \lambda_0, \lambda_1, \mu) = 0 \iff I(P, \lambda_0, \lambda_1, 1) = 0 \tag{40}$$

We will prove in appendix B that

Theorem 4.2: For any probability matrices P_1, P_2 and their tensor product $P_1 \times P_2$

$$I(P_1 \times P_2, \lambda_0, \lambda_1, \mu) = I(P_1, \lambda_0, \lambda_1, \mu) + I(P_2, \lambda_0, \lambda_1, \mu) \tag{41}$$

Because (34) if opaque, we will use the longer but easier to manipulate definition A.1. Their equivalence will be proven for $\mu = 1$ in section VII. The general equivalence is not used in this paper, so is relegated to [6].

V. Main results

Our key lower bound is

Theorem 5.1: For any d dimensional bucketing code handling data with probability matrices $P_1, \ldots, P_d$, set sizes n_0, n_1, success probability S and work W

$$\ln W \ge \sup_{0\le\lambda_0,\lambda_1\le 1\le\mu,\lambda_0+\lambda_1}\left[\lambda_0\ln n_0 + \lambda_1\ln n_1 + \mu\ln S - \sum_{i=1}^{d} I(P_i, \lambda_0, \lambda_1, \mu)\right] \tag{42}$$

This will be proven in appendix C. In appendix D we will show that the lower bound is asymptotically tight:

Definition 5.1: For any probability matrix P let

$$\omega(P) = -\ln p_{j_1*}p_{*k_1} \tag{43}$$

where indices j_1, k_1 maximize $\frac{p_{jk}}{p_{j*}p_{*k}}$.

Theorem 5.2: For any $\epsilon > 0$ there exits a constant $a(\epsilon) > 0$ such that for any dimension $d \ge 1$, probabilities satisfying $p_{i,jk} = 0$ or $p_{i,jk} > \epsilon$ for any $1 \le i \le d$, $0 \le j < b_0$, $0 \le k < b_1$ and parameters $1 \le n_0, n_1 \le W$, $0 < S \le 1$ satisfying (42) and $W \ge n_0 n_1 e^{-\sum_i \omega(P_i)}$ there exists a bucketing code with work $\tilde{W} \le W\mathrm{e}^{a(\epsilon)+\epsilon d}$ vector operations and success probability $\tilde{S} \ge S\mathrm{e}^{-a(\epsilon)-\epsilon d}$. The memory requirements are $a(\epsilon)$ size code table and $n_0 + n_1$ vectors of pointers.

The bound $W \ge n_0 n_1 e^{-\sum_i \omega(P_i)}$ has effect beyond $W \ge n_0, n_1$ only when $\min(n_0, n_1)e^{-\sum_i \omega(P_i)} > 1$, which implies both $n_0\prod_{i=1}^d p_{i,j_i*} > 1$ and $n_1\prod_{i=1}^d p_{i,*k_i} > 1$ i.e. we expect several data points to be indistinguishable. This could happen.

For example when $d = 0$ we can either reject all pairs: $S = 0, W = 0$ or accept all pairs: $S = 1, W = n_0 n_1$.

The loss factor $e^{-\epsilon d}$ seems excessive. Even $\tilde{S} \geq 1/2$ might be considered too low. For constant $P_i = P$ it is indeed possible to utilize the approach of section II to obtain $\tilde{S} \geq S(1 - o(1))$, or even $\tilde{S} \geq S(1 - e^{-\epsilon d})$ by replacing Chebyshev's inequality with Chernoff's inequality. We are more interested in making a first stab at the much neglected heterogeneous problem.

An equivalent way of writing condition (42) is

$$(\ln n_0, \ln n_1, -\ln S, \ln W) \in \sum_{i=1}^{d} D(P_i) \tag{44}$$

where

Definition 5.2: For any probability matrix P, its $\mathbf{log - attainable}$ set is

$$D(P) = \{(m_0, m_1, s, w) \mid \forall\ \lambda_0, \lambda_1 \leq 1 \leq \mu, \lambda_0 + \lambda_1 \\ w \geq \lambda_0 m_0 + \lambda_1 m_1 - \mu s - I(P, \lambda_0, \lambda_1, \mu)\}\Big] \tag{45}$$

In terms of log-attainable sets, theorem 4.2 is equivalent to

$$D(P_1 \times P_2) = D(P_1) + D(P_2) \tag{46}$$

Intuitively the meaning of the critical λ_0 (attaining the supremum) is that when we double n_0, the work increases approximately by a factor of 2^{λ_0}. The reason for $\lambda_0 \geq 0$ is that increasing n_0 does not decrease the work. The reason for $\lambda_0 \leq 1$ is that doubling n_0 can at most double the work (use the same bucketing code). Similarly $0 \leq \lambda_1 \leq 1$. The reason for $\lambda_0 + \lambda_1 \geq 1$ is that if we double n_0 and n_1 the work must be at least doubled. The meaning of the critical μ is that when we double the work, the success probability increases approximately by a factor of $2^{1/\mu}$. The reason for $\mu \geq 1$ is that doubling the work at most doubles the success probability (the larger S is, the harder it is to improve).

In order to illustrate how (42) works let us consider diagonal P ($p_{jk} = 0$ for $j \neq k$). There are no errors, so the best codes must consist of pairs of identical points. When we demand $S = 1$ the supremum of (42) is attained by $\mu = \infty$. For a diagonal matrix $I(P, \lambda_0, \lambda_1, \infty) = (\lambda_0 + \lambda_1 - 1)I(P)$ where $I(P) = -\sum_{j=0}^{b-1} p_{jj} \ln p_{jj}$ is the information. Insertion into (42) reveals the expected $\ln W \geq \ln n_0 + \ln n_1 - \sum_{i=1}^{d} I(P_i)$. Even when $S < 1$ and $\mu < \infty$ for a diagonal P, the function $I(P, \lambda_0, \lambda_1, \mu)$ turns out to be a familiar object of large deviation theory.

When there are errors we can not insist on a perfect $S = 1$. Inserting $\lambda_0 = \lambda_1 = 1,\ \mu = \ln\ln(n_0 + n_1)$ into (42) reveals that

$$\ln W \geq \ln n_0 + \ln n_1 + \ln\ln(n_0 + n_1) \ln S - \\ - \sum_{i=1}^{d} I\left(P, 1, 1, \ln\ln(n_0 + n_1)\right) \tag{47}$$

For $n_0 + n_1 \to \infty$ $I\left(P, 1, 1, \ln\ln(n_0 + n_1)\right) \to I(P, 1, 1, \infty)$ which turns out to be the familiar mutual information $I(P)$, so we have recreated a version of the information theoretic (2).

Let us reconsider $P_i = \begin{pmatrix} p/2 & (1-p)/2 \\ (1-p)/2 & p/2 \end{pmatrix}$, $n_0 = n_1 = n$ from section II. The matrix P is symmetric so

$$I(P, \lambda_0, \lambda_1, \mu) = I(P, \lambda_1, \lambda_0, \mu) \tag{48}$$

The convexity of I implies

$$I(P, \lambda_0, \lambda_1, \mu) \geq I(P, (\lambda_0 + \lambda_1)/2, (\lambda_0 + \lambda_1)/2, \mu) \tag{49}$$

so in theorems 5.1,5.2 it is enough to consider $\lambda_0 = \lambda_1 = \lambda$. Definition A.1 involves a finite dimensional maximum, so it is easy to obtain lower bounds. We will show in section VIII that for any $\epsilon > 0$

$$I(P, 1/(2p) + \epsilon, 1/(2p) + \epsilon, \infty) > 0 \tag{50}$$

Hence for any $1/2 \leq \lambda \leq 1 \leq \mu$, $d \geq \ln n / I(P, 1/(2p) + \epsilon, 1/(2p) + \epsilon, \infty)$

$$2\lambda \ln n + \mu \ln S - I(P, \lambda, \lambda, \mu) d \leq \\ \leq 2\lambda \ln n - I(P, \lambda, \lambda, \infty) d \leq (1/p + 2\epsilon) \ln n \tag{51}$$

(consider $\lambda \leq 1/(2p) + \epsilon$ and $\lambda \geq 1/(2p) + \epsilon$ separately). Theorem 5.2 says that with work $W \leq n^{1/p + 3\epsilon} \mathrm{e}^{a(\epsilon)}$ we can attain success probability $n^{-\epsilon} \mathrm{e}^{-a(\epsilon)}$. In the other direction

$$\ln W \geq 1/p \ln n + \ln S - I(P, 1/(2p), 1/(2p), 1) d \tag{52}$$

In a separate paper [6] we will prove

Theorem 5.3:

$$I\left(\begin{pmatrix} p/2 & (1-p)/2 \\ (1-p)/2 & p/2 \end{pmatrix}, 1/(2p), 1/(2p), 1\right) = 0 \tag{53}$$

Hence $1/p$ is indeed the critical exponent.

Thus we have obtained a near tight lower bound involving a generalization of mutual information, which is asymptotically approximated by a randomly constructed block code. The analogy with Shannon's coding and information theory (and coding with distortion theory) is very strong, suggesting that maybe we are redoing it in disguise. If it is a disguise, it is quite effective. At least we feel fully justified calling $I(P, \lambda_0, \lambda_1, \mu)$ the $\mathbf{bucketing\ information}$ function.

We also refer the reader to [5], which tackles a particular class of practical bucketing algorithms (small leaves bucketing forests). Their performance turns out to be bounded by a $\mathbf{small\ leaves\ bucketing\ information}$ function, and that bound is asymptotically attained by a specific practical algorithm.

VI. Implication for the Indyk-Motwani Theory

The Indyk-Motwani paper [8] introduces a metric based, worst case analysis. It is (nonessentially) monochromatic, meaning that $X_0 = X_1 = X \subset \mathrm{R}^d$, $n_0 = n_1 = n$. There is a metric dist and constants $c \geq 1$, $r > 0$ such that matched pairs $x, y \in X$ satisfy $\mathrm{dist}(x, y) \leq r$ while random pairs satisfy $\mathrm{dist}(x, y) \geq cr$. They have shown that for an L_1 metric $\mathrm{dist}(x, y) = \sum_{i=1}^{d} |x_i - y_i|$ there exists an algorithm (called LSH) with success probability $S = 1 - o(1)$ and work $W = O(n^{1+1/c})$. We prefer Shannon's information theoretic formulation, for reasons explained in [5]. Nevertheless an LSH is a bucketing code obeying certain

symmetry and disjointness requirements, so our lower work bounds apply. Let the metric dist^s be the standard L_s distance and $P_i = \begin{pmatrix} p/2 & (1-p)/2 \\ (1-p)/2 & p/2 \end{pmatrix}$, $d \to \infty$. By the law of large numbers for a matched pair (x,y) $\mathrm{dist}^s(x,y) = (1-p+o(1))d$, while for a random pair $\mathrm{dist}^s(x,y) = (1/2+o(1))d$ so $c^s = 1/(2-2p)+o(1)$. Theorem 5.3 implies that in order to achieve success probability $1/2$ the LSH will have at least $W \geq n^{1/p}/2 = n^{1+1/(2c^s-1)+o(1)}$. In [8]'s terms

$$\rho_s(c) \geq 1/(2c^s-1) \tag{54}$$

which slightly improves Motwani Naor and Panigrahy's [11] $\rho_s(c) \geq \frac{\mathrm{e}^{c^{-s}}-1}{\mathrm{e}^{c^{-s}}+1}$.

VII. A PROOF FROM THE BOOK

The general proofs are quite complicated. However there is an interesting simpler special case: $\mu = 1$. Let us see when it comes into play. Suppose we have unlimited homogeneous data $P_i = P$ for $i = 1, 2, \ldots$. What do **theorems 5.1,5.2** say? Whenever $I(P, \lambda_0, \lambda_1, \mu) > 0$ we can take d large enough so that the bound means nothing. So the only distinction is whether $I(P, \lambda_0, \lambda_1, \mu) = 0$ or not. Because of (40) this is independent of μ, so $\mu = 1$ gives the stronger bounds. In short

$$\ln W \geq \sup_{\substack{\lambda_0,\lambda_1 \leq 1 \leq \lambda_0+\lambda_1 \\ I(P,\lambda_0,\lambda_1,1)=0}} [\lambda_0 \ln n_0 + \lambda_1 \ln n_1 + \ln S] \tag{55}$$

and this is asymptotically tight. It can be written as

$$(\ln n_0, \ln n_1, -\ln S, \ln W) \in \mathrm{Cone}(D(P)) \tag{56}$$

where we define the **log-attainable cone**

$$\begin{aligned} &\mathrm{Cone}(D(P)) = \cup_{\alpha \geq 0} \alpha D(P) = \{(m_0, m_1, s, w) \mid \\ &\quad \forall\ \lambda_0, \lambda_1 \leq 1 \leq \lambda_0 + \lambda_1 \\ &I(P, \lambda_0, \lambda_1, 1) = 0 \implies w \geq \lambda_0 m_0 + \lambda_1 m_1 - s\} \end{aligned} \tag{57}$$

The constructive part of the proof (asymptotic tightness) is similar to what we did in section II. In this section we will prove the lower bound.

We begin with some definitions.

Definition 7.1: For any nonnegative matrix or vector R, and a probability matrix or vector P of the same dimensions $b_0 \times b_1$, let the extended Kullback-Leibler divergence be

$$K(R\|P) = \sum_{j=0}^{b_0-1} \sum_{k=0}^{b_1-1} r_{jk} \ln \frac{r_{jk}}{r_{**} p_{jk}} \geq 0 \tag{58}$$

where $r_{**} = \sum_{j=0}^{b_0-1} \sum_{k=0}^{b_1-1} r_{jk}$

Non-negativity follows from the well known log-sum inequality:

Lemma 7.1: For any nonnegative $q_0, q_1, \ldots, q_{b-1} \geq 0$, $p_0, p_1, \ldots, p_{b-1} \geq 0$

$$\sum_{j=0}^{b-1} q_j \ln \frac{q_j}{p_j} \geq q_* \ln \frac{q_*}{p_*} \tag{59}$$

where $q_* = \sum_{j=0}^{b-1} q_j$, $p_* = \sum_{j=0}^{b-1} p_j$

The bucketing information function for $\mu = 1$ is defined by

Definition 7.2: Suppose P is a probability matrix. For any $\lambda_0, \lambda_1 \leq 1 \leq \lambda_0 + \lambda_1$

$$\begin{aligned} I(P, \lambda_0, \lambda_1, 1) = \max_Q [&\lambda_0 K(Q_{\cdot *}\|P_{\cdot *}) + \\ &+ \lambda_1 K(Q_{* \cdot}\|P_{* \cdot}) - K(Q_{\cdot\cdot}\|P_{\cdot\cdot})] \end{aligned} \tag{60}$$

where Q ranges over all probability matrices of the same dimensions as P.

Explicitly

$$\begin{aligned} I(P, \lambda_0, \lambda_1, 1) = \max_{\substack{\{q_{jk} \geq 0\}_{jk} \\ 0 \leq j < b_0 \\ 0 \leq k < b_1 \\ q_{**}=1}} \Bigg[&\lambda_0 \sum_{j=0}^{b_0-1} q_{j*} \ln \frac{q_{j*}}{p_{j*}} + \\ &+ \lambda_1 \sum_{k=0}^{b_1-1} q_{*k} \ln \frac{q_{*k}}{p_{*k}} - \sum_{j=0}^{b_0-1} \sum_{k=0}^{b_1-1} q_{jk} \ln \frac{q_{jk}}{p_{jk}} \Bigg] \end{aligned} \tag{61}$$

This definition makes it clear that the bucketing information function is nonnegative (insert $Q = P$), monotonically non-decreasing in λ_0, λ_1 (lemma 7.1) and is convex in λ_0, λ_1 (maximum of linear functions).

Let us reconcile with (34):

Theorem 7.2:

$$\begin{aligned} &I(P, \lambda_0, \lambda_1, 1) = \\ &= \max_{\substack{\{x_{0,j} \geq 0\}_j \\ 0 \leq j < b_0 \\ x_{0,*}=1}} \max_{\substack{\{x_{1,k} \geq 0\}_k \\ 0 \leq k < b_1 \\ x_{1,*}=1}} \ln \sum_{j=0}^{b_0-1} \sum_{k=0}^{b_1-1} p_{jk} \left(\frac{x_{0,j}}{p_{j*}}\right)^{\lambda_0} \left(\frac{x_{1,k}}{p_{*k}}\right)^{\lambda_1} = \end{aligned} \tag{62}$$

$$= \max_{\substack{\{x_{1,k} \geq 0\}_k \\ 0 \leq k < b_1 \\ x_{1,*}=1}} (1-\lambda_0) \ln \sum_{j=0}^{b_0-1} \left[\sum_{k=0}^{b_1-1} p_{jk} p_{j*}^{-\lambda_0} \left(\frac{x_{1,k}}{p_{*k}}\right)^{\lambda_1} \right]^{\frac{1}{1-\lambda_0}} \tag{63}$$

Proof: Using 7.1, (61) can be rewritten as

$$\begin{aligned} I(P, \lambda_0, \lambda_1, 1) = &\max_{\substack{\{q_{jk} \geq 0\}_{jk} \\ 0 \leq j < b_0 \\ 0 \leq k < b_1 \\ q_{**}=1}} \max_{\substack{\{x_{0,j} \geq 0\}_j \\ 0 \leq j < b_0 \\ x_{0,*}=1}} \max_{\substack{\{x_{1,k} \geq 0\}_k \\ 0 \leq k < b_1 \\ x_{1,*}=1}} \\ &\sum_{jk} q_{jk} \left[\lambda_0 \ln \frac{x_{0,j}}{p_{j*}} + \lambda_1 \ln \frac{x_{1,k}}{p_{*k}} - \ln \frac{q_{jk}}{p_{jk}} \right] \end{aligned} \tag{64}$$

Again 7.1 provides the optimal q_{jk}'s

$$q_{jk} = p_{jk} \left(\frac{x_{0,j}}{p_{j*}}\right)^{\lambda_0} \left(\frac{x_{1,k}}{p_{*k}}\right)^{\lambda_1} \Bigg/ \sum_{\tilde{j}\tilde{k}} p_{\tilde{j}\tilde{k}} \left(\frac{x_{0,\tilde{j}}}{p_{\tilde{j}*}}\right)^{\lambda_0} \left(\frac{x_{1,\tilde{k}}}{p_{*\tilde{k}}}\right)^{\lambda_1} \tag{65}$$

Insertion into (64) gives (62). Then standard $\mathrm{L}_{\frac{1}{\lambda_0}}$, $\mathrm{L}_{\frac{1}{1-\lambda_0}}$ duality proves (63). ∎

Proof of theorem 4.2 for $\mu = 1$:

$$I(P_1 \times P_2, \lambda_0, \lambda_1, 1) = I(P_1, \lambda_0, \lambda_1, 1) + I(P_2, \lambda_0, \lambda_1, 1) \tag{66}$$

Proof: Inequality $I(P_1 \times P_2, \lambda_0, \lambda_1, 1) \geq I(P_1, \lambda_0, \lambda_1, 1) + I(P_2, \lambda_0, \lambda_1, 1)$ is obvious because we can choose $Q = Q_1 \times Q_2$. The other direction is the

challenge. Denote $I_1 = I(P_1, \lambda_0, \lambda_1, 1)$, $I_2 = I(P_2, \lambda_0, \lambda_1, 1)$ and $P = P_1 \times P_2$

$$p_{j_1k_1j_2k_2} = p_{1,j_1k_1}p_{2,j_2k_2} \tag{67}$$

The nonstandard $j_1k_1j_2k_2$ ordering of coordinates will be used throughout this proof. For any probability matrix $\{q_{j_1k_1j_2k_2}\}_{j_1k_1j_2k_2}$

$$\sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{p_{1,j_1k_1}p_{2,j_2k_2}} =$$
$$= \sum_{j_1k_1} q_{j_1k_1**} \ln \frac{q_{j_1k_1**}}{p_{1,j_1k_1}} + \sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{q_{j_1k_1**}p_{2,j_2k_2}} \tag{68}$$

By definition

$$I_1 + \sum_{j_1k_1} q_{j_1k_1**} \ln \frac{q_{j_1k_1**}}{p_{1,j_1k_1}} \geq$$
$$\geq \lambda_0 \sum_{j_1} q_{j_1***} \ln \frac{q_{j_1***}}{p_{1,j_1*}} + \lambda_1 \sum_{k_1} q_{*k_1**} \ln \frac{q_{*k_1**}}{p_{1,*k_1}} \tag{69}$$

and for all j_1k_1

$$I_2 + \sum_{j_2k_2} q_{j_1k_1j_2k_2}/q_{j_1k_1**} \ln \frac{q_{j_1k_1j_2k_2}/q_{j_1k_1**}}{p_{2,j_2k_2}} \geq$$
$$\geq \lambda_0 \sum_{j_2} q_{j_1k_1j_2*}/q_{j_1k_1**} \ln \frac{q_{j_1k_1j_2*}/q_{j_1k_1**}}{p_{2,j_2*}} +$$
$$\lambda_1 \sum_{k_2} q_{j_1k_1*k_2}/q_{j_1k_1**} \ln \frac{q_{j_1k_1*k_2}/q_{j_1k_1**}}{p_{2,*k_2}} \tag{70}$$

Thus when multiplying by $q_{j_1k_1**}$ and summing over j_1k_1

$$I_2 + \sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{q_{j_1k_1**}p_{2,j_2k_2}} \geq$$
$$\geq \lambda_0 \sum_{j_1k_1j_2} q_{j_1k_1j_2*} \ln \frac{q_{j_1k_1j_2*}}{q_{j_1k_1**}p_{2,j_2*}} +$$
$$+\lambda_1 \sum_{j_1k_1k_2} q_{j_1k_1*k_2} \ln \frac{q_{j_1k_1*k_2}}{q_{j_1k_1**}p_{2,*k_2}} \tag{71}$$

so with help from lemma 7.1

$$I_2 + \sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{q_{j_1k_1**}p_{2,j_2k_2}} \geq$$
$$\geq \lambda_0 \sum_{j_1j_2} q_{j_1*j_2*} \ln \frac{q_{j_1*j_2*}}{q_{j_1***}p_{2,j_2*}} +$$
$$+\lambda_1 \sum_{k_1k_2} q_{*k_1*k_2} \ln \frac{q_{*k_1*k_2}}{q_{*k_1**}p_{2,*k_2}} \tag{72}$$

Combining formulas (68),(69) and (72) gives

$$I_1 + I_2 + \sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{p_{1,j_1k_1}p_{2,j_2k_2}} \geq$$
$$\geq \lambda_0 \sum_{j_1j_2} q_{j_1*j_2*} \ln \frac{q_{j_1*j_2*}}{p_{1,j_1*}p_{2,j_2*}} +$$
$$+\lambda_1 \sum_{k_1k_2} q_{*k_1*k_2} \ln \frac{q_{*k_1*k_2}}{p_{1,*k_1}p_{2,*k_2}} \tag{73}$$

hence $I_1 + I_2 \geq I(P_1 \times P_2, \lambda_0, \lambda_1, 1)$. ■

Theorem 7.3: For any $B_0 \subset \{0, 1, \ldots, b_0 - 1\}^d$, $B_1 \subset \{0, 1, \ldots, b_1 - 1\}^d$

$$p_{B_0B_1} \leq \min_{\lambda_0,\lambda_1 \leq 1 \leq \lambda_0+\lambda_1} p_{B_0*}^{\lambda_0} p_{*B_1}^{\lambda_1} e^{I(P,\lambda_0,\lambda_1,1)d} \tag{74}$$

Proof: Without restricting generality let $d = 1$. Inserting

$$q_{jk} = \begin{cases} \frac{p_{jk}}{p_{B_0B_1}} & j \in B_0, k \in B_1 \\ 0 & \text{otherwise} \end{cases} \tag{75}$$

into (61) and using lemma 7.1 proves the assertion. ■

We can now prove an important special case of theorem 5.1:

Theorem 7.4: For any bucketing code with probability matrices $P_i = P$, set sizes n_0, n_1, success probability S and work W

$$W \geq S \sup_{\lambda_0,\lambda_1 \leq 1 \leq \lambda_0+\lambda_1} n_0^{\lambda_0} n_1^{\lambda_1} e^{-I(P,\lambda_0,\lambda_1,1)d} \tag{76}$$

Proof: Without restricting generality let $d = 1$. By definition $3W \geq \sum_t W_t$ where

$$W_t = \max\left(n_0 p_{B_{0,t}*},\ n_1 p_{*B_{1,t}},\ n_0 p_{B_{0,t}*} n_1 p_{*B_{1,t}}\right) \tag{77}$$

so for $(\lambda_0, \lambda_1) = (0, 1), (1, 0), (1, 1)$ we already have

$$\ln W_t \geq \lambda_0 \ln(n_0 p_{B_{0,t}*}) + \lambda_1 \ln(n_1 p_{*B_{1,t}}) \tag{78}$$

and by linearity it holds for all $(\lambda_0, \lambda_1) \in \mathrm{Conv}(\{(1, 0), (0, 1), (1, 1)\})$. With the help of (74)

$$W_t \geq (n_0 p_{B_{0,t}*})^{\lambda_0} (n_1 p_{*B_{1,t}})^{\lambda_1} \geq$$
$$\geq n_0^{\lambda_0} n_1^{\lambda_1} p_{B_{0,t}B_{1,t}} e^{-I(P,\lambda_0,\lambda_1,1)} \tag{79}$$

Summing up over t proves

$$3W \geq n_0^{\lambda_0} n_1^{\lambda_1} S e^{-I(P,\lambda_0,\lambda_1,1)} \tag{80}$$

In order to get rid of the pesky 3 we concatenate the bucketing code with itself $d \to \infty$ times. Then

$$3W^d \geq n_0^{\lambda_0 d} n_1^{\lambda_1 d} S^d e^{-I(P,\lambda_0,\lambda_1,1)d} \tag{81}$$

so taking d'th root concludes this proof. ■

VIII. Computing Log-Attainable Points

Again $P = \begin{pmatrix} p/2 & (1-p)/2 \\ (1-p)/2 & p/2 \end{pmatrix}$, $d \to \infty$. Inserting $Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ into (60) (or $x_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ into (63)) implies that $I(P, \lambda_0, \lambda_1, 1) \geq p2^{\lambda_0+\lambda_1-1}$. Hence $\lambda_0 + \lambda_1 > \log_2 2/p \Longrightarrow I(P, \lambda_0, \lambda_1, 1) > 0$. Recalling (57), it is the long known

$$(1,\ 1,\ 0,\ \log_2 2/p) \in \mathrm{Cone}(D(P)) \tag{82}$$

Formula (63) for our specific matrix is

$$e^{I(P,\lambda_0,\lambda_1,1)} = \frac{1}{2} \max_{0 \leq x \leq 1} \left[f(x) + f(1-x)\right] \tag{83}$$

where

$$f(x) = \left[p(2x)^{\lambda_1} + (1-p)(2-2x)^{\lambda_1}\right]^{\frac{1}{1-\lambda_0}} \tag{84}$$

Because $f(1/2) = 1$

$$\frac{d^2 f(1/2)}{dx^2} > 0 \implies I(P, \lambda_0, \lambda_1, 1) > 0 \tag{85}$$

Differentiation gives

$$\frac{(1-\lambda_0)^2}{4\lambda_1}\frac{d^2f(1/2)}{dx^2} = (2p-1)^2\lambda_0\lambda_1 - (1-\lambda_0)(1-\lambda_1) \quad (86)$$

In particular $\lambda_0+\lambda_1 > 1/p \implies I(P,\lambda_0,\lambda_1,1) > 0$ so

$$(1,\ 1,\ 0,\ 1/p) \in \mathrm{Cone}(D(P)) \quad (87)$$

which is essentially section II. We have found a proof [6] that $I(P,1/2p,1/2p,1)=0$ so (87) is right on the boundary.

Another interesting log-attainable point is found by looking near $(\lambda_0,\lambda_1) = (0,1)$: $(2p-1)^2\lambda_0+\lambda_1 > 1 \implies I(P,\lambda_0,\lambda_1,1) > 0$ so

$$((2p-1)^2,\ 1,\ 0,\ 1) \in \mathrm{Cone}(D(P)) \quad (88)$$

This means that when $n_0 \le n_1^{(2p-1)^2}$ we can find matched pairs in near linear time!

IX. Conclusion

We consider the approximate nearest neighbor problem in a probabilistic setting. Using several coordinates at once enables asymptotically better approximate nearest neighbor algorithms than using them one at a time. The performance is bounded by, and tends to, a newly defined bucketing information function. Thus bucketing coding and information theory play the same role for the approximate nearest neighbor problem that Shannon's coding and information theory play for communication.

Appendix A
Bucketing Information Basics

Definition A.1: Suppose P is a probability matrix. The **bucketing information** function is for $\lambda_0,\lambda_1 \le 1 \le \mu, \lambda_0+\lambda_1$

$$I(P,\lambda_0,\lambda_1,\mu) = \max_{\substack{\{r_{i,jk}\ge 0\}_{ijk}\\ 0\le i<b_0b_1\\ 0\le j<b_0\\ 0\le k<b_1\\ r_{*,**}=1}} \Bigg[\lambda_0 \sum_{i=0}^{b_0b_1-1} K(R_{i,\cdot *}\|P_{*}) + \lambda_1 \sum_{i=0}^{b_0b_1-1} K(R_{i,*\cdot}\|P_{*\cdot}) - (\mu-1)K(R_{*,\cdot\cdot}\|P_{\cdot\cdot}) - \sum_{i=0}^{b_0b_1-1} K(R_{i,\cdot\cdot}\|P_{\cdot\cdot})\Bigg] \quad (89)$$

Explicitly $r_{i,j*} = \sum_{k=0}^{b_1-1} r_{i,jk}$, $K(R_{i,\cdot *}\|P_{*}) = \sum_{j=0}^{b_0-1} r_{i,j*}\ln\frac{r_{i,j*}}{r_{i,**}p_{j*}}$ etc.
The quantity maximized upon looks a bit like curvature. Insertion into definition 5.2 results in

Lemma A.1:

$$D(P) = \mathrm{Conv}^c\Bigg(\Big\{\Big(\sum_i K(R_{i,\cdot *}\|P_{*}),\ \sum_i K(R_{i,*\cdot}\|P_{*\cdot}),\ K(R_{*,\cdot\cdot}\|P_{\cdot\cdot}),\ -K(R_{*,\cdot\cdot}\|P_{\cdot\cdot}) + \sum_i K(R_{i,\cdot\cdot}\|P_{\cdot\cdot})\Big)\Big\}_R\Bigg) + \mathrm{ConvCone}^c(\{(1,0,0,1),(0,1,0,1),(-1,-1,0,-1), ,(0,0,1,0),(0,0,0,1),(0,0,1,-1)\}) \quad (90)$$

where Conv^c is the closure of the convex hull and $\mathrm{ConvCone}^c$ is the closure of the convex cone (nonnegative span). In particular

$$\mathrm{Cone}(D(P)) = \mathrm{ConvCone}^c(\{(K(Q_{\cdot *}\|P_{*}), , K(Q_{*\cdot}\|P_{*\cdot}),\ 0,\ K(Q_{\cdot\cdot}\|P_{\cdot\cdot}))\}_Q) + \mathrm{ConvCone}(\{(1,0,0,1),(0,1,0,1),(-1,-1,0,-1), ,(0,0,1,0),\pm(0,0,1,-1)\}) \quad (91)$$

where Q runs over all $b_0\times b_1$ probability matrices.

Proof: We will prove (91). By linear duality inequalities $\lambda_0\cdot 1+\lambda_1\cdot 0 \le 1$, $\lambda_0\cdot 0+\lambda_1\cdot 1 \le 1$, $\lambda_0\cdot(-1)+\lambda_1\cdot(-1) \le -1$ and $\lambda_0\cdot K(Q_{\cdot *}\|P_{*})+\lambda_1\cdot K(Q_{*\cdot}\|P_{*\cdot}) \le K(Q_{\cdot\cdot}\|P_{\cdot\cdot})$ (for several Q's) imply $\lambda_0\cdot m_0+\lambda_1\cdot m_1 \le s+w$ if $(m_0,m_1,s+w)$ is a nonnegative combination of $(1,0,1)$, $(0,1,1)$, $(-1,-1,-1)$, $(0,0,1)$ and several $(K(Q_{\cdot *}\|P_{*}),K(Q_{*\cdot}\|P_{*\cdot}),K(Q_{\cdot\cdot}\|P_{\cdot\cdot}))$. The proof of (90) is very similar. ∎

Let us take a closer look at (89). Not restricting the number of terms i does not change I. It can be rewritten as:

Lemma A.2:

$$I(P,\lambda_0,\lambda_1,\mu) = \max_Q \Bigg[-(\mu-1)K(Q_{\cdot\cdot}\|P_{\cdot\cdot}) + \max_{(Q,y)\in \mathrm{Conv}^c(\Delta(P,\lambda_0,\lambda_1))} y\Bigg] \quad (92)$$

where

$$\Delta(P,\lambda_0,\lambda_1) = \{(Q,\ \lambda_0 K(Q_{\cdot *}\|P_{*}) + \lambda_1 K(Q_{*\cdot}\|P_{*\cdot}) - K(Q_{\cdot\cdot}\|P_{\cdot\cdot}))\}_Q \quad (93)$$

Proof: The connection between definition A.1 and (92) is through $r_i = r_{i,**}$, $q_{i,jk} = \frac{r_{i,jk}}{r_{i,**}}$

$$I(P,\lambda_0,\lambda_1,\mu) = \max_{\substack{\{r_i,Q_i\}_i\\ r_*=1}} \Bigg[\sum_{i=0}^{b_0b_1-1} r_i\Big[\lambda_0 K(Q_{i,\cdot *}\|P_{*}) + \lambda_1 K(Q_{i,*\cdot}\|P_{*\cdot}) - (\mu-1)K\Big(\sum_i r_i Q_{i,\cdot\cdot}\Big\|P_{\cdot\cdot}\Big) - K(Q_{i,\cdot\cdot}\|P_{\cdot\cdot})\Big]\Bigg] \quad (94)$$

The set Δ is b_0b_1 dimensional, so by Caratheodory's theorem any point on the boundary of its convex hull is a convex combination of b_0b_1 points from Δ. ∎

From now on when dealing with the bucketing information function, we will write $\sum_i$ without worrying about the number of indices.

Lemma A.3: For any probability matrix P the bucketing information function $I(P,\lambda_0,\lambda_1,\mu)$ is nonnegative, monotonically nondecreasing and convex in $\lambda_0,\lambda_1,-\mu$. Special values are

$$I(P,\lambda_0,\lambda_1,\mu) = 0 \iff \iff \forall Q,\ K(Q_{\cdot\cdot}\|P_{\cdot\cdot}) \ge \lambda_0 K(Q_{\cdot *}\|P_{*}) + \lambda_1 K(Q_{*\cdot}\|P_{*\cdot}) \quad (95)$$

$$I(P,\lambda_0,1-\lambda_0,\mu) = 0 \quad (96)$$

$$I(P,1,1,\mu) = (\mu-1)\ln\sum_{j=0}^{b_0-1}\sum_{k=0}^{b_1-1} p_{jk}\left(\frac{p_{jk}}{p_{j*}p_{*k}}\right)^{\frac{1}{\mu-1}} \quad (97)$$

$$I(P,1,1,1)=\ln\max_{\substack{0\le j<b_0\\0\le k<b_1}}\frac{p_{jk}}{p_{j*}p_{*k}} \tag{98}$$

$$I(P,1,1,\infty)=I(P)=\sum_{j=0}^{b_0-1}\sum_{k=0}^{b_1-1}p_{jk}\ln\frac{p_{jk}}{p_{j*}p_{*k}} \tag{99}$$

For a diagonal P ($p_{jk}=0$ for $j\neq k$)

$$I(P,\lambda_0,\lambda_1,\mu)=(\mu-1)\ln\sum_{j=0}^{b-1}p_{jj}^{\frac{\mu-\lambda_0-\lambda_1}{\mu-1}} \tag{100}$$

Proof: Non-negativity follows by taking $Q=P$. Monotonicity and convexity hold by definition. Clearly

$$I(P,\lambda_0,\lambda_1,\mu)\le\max_Q\sum_{i=0}^{b_0b_1-1}\Big[\lambda_0K(Q_{i,\cdot *}\|P_{\cdot *})+ \\ +\lambda_1K(Q_{i,*\cdot}\|P_{*\cdot})-K(Q_{i,\cdot\cdot}\|P_{\cdot\cdot})\Big] \tag{101}$$

so direction $\Leftarrow$ of (95) is true. On the other hand assume that for some Q

$$K(Q_{\cdot\cdot}\|P_{\cdot\cdot})<\lambda_0K(Q_{\cdot *}\|P_{\cdot *})+\lambda_1K(Q_{*\cdot}\|P_{*\cdot}) \tag{102}$$

Inserting $r_{0,jk}=\epsilon q_{jk},\ r_{1,jk}=p_{jk}-\epsilon q_{jk}$ into definition A.1 gives

$$I(P,\lambda_0,\lambda_1,\mu)\ge\epsilon\Big[\lambda_0K(Q_{\cdot *}\|P_{\cdot *})+ \\ +\lambda_1K(Q_{*\cdot}\|P_{*\cdot})-K(Q_{\cdot\cdot}\|P_{\cdot\cdot})\Big]+ \\ +(1-\epsilon)\Big[\lambda_0K(\tilde{P}_{\cdot *}\|P_{\cdot *})+\lambda_1K(\tilde{P}_{*\cdot}\|P_{*\cdot})-K(\tilde{P}_{\cdot\cdot}\|P_{\cdot\cdot})\Big] \tag{103}$$

where $\tilde{P}=(P-\epsilon Q)/(1-\epsilon)=P+\epsilon(P-Q)/(1-\epsilon)$. The Kullback-Leibler divergence between $\tilde{P}$ and P is second order in ϵ, and the same holds for their marginal vectors. Hence for a small $\epsilon>0$ the bucketing information $I(P,\lambda_0,\lambda_1,\mu)$ is positive, and the proof of (95) is done.

Lemma 7.1 implies

$$K(R_{i,\cdot *}\|P_{\cdot *}),K(R_{i,*\cdot}\|P_{*\cdot})\le K(R_{i,\cdot\cdot}\|P_{\cdot\cdot}) \tag{104}$$

so (96) follows from (95).

Now will prove (97). We want to maximize

$$\sum_i\big[K(R_{i,\cdot *}\|P_{\cdot *})+K(R_{i,*\cdot}\|P_{*\cdot})-K(R_{i,\cdot\cdot}\|P_{\cdot\cdot})\big]= \\ =\sum_{jk}r_{*,jk}\ln\frac{p_{jk}}{p_{j*}p_{*k}}-\sum_{ijk}r_{i,jk}\ln\frac{r_{i,**}r_{i,jk}}{r_{i,j*}r_{i,*k}} \tag{105}$$

The rightmost sum is nonnegative, and for any $\{r_{*,jk}\}_{jk}$ it can be made 0 by choosing

$$r_{i,jk}=\begin{cases}r_{*,jk} & i=j+b_0k\\ 0 & \text{otherwise}\end{cases} \tag{106}$$

Hence we want to maximize

$$\sum_{jk}r_{*,jk}\ln\frac{(p_{jk})^\mu}{p_{j*}p_{*k}}+(1-\mu)\sum_{jk}r_{*,jk}\ln r_{*,jk} \tag{107}$$

The maximized function is concave in $\{r_{*,jk}\}_{jk}$, and Lagrange multipliers reveal the optimal choice

$$r_{*,jk}=\left(\frac{(p_{jk})^\mu}{p_{j*}p_{*k}}\right)^{\frac{1}{\mu-1}}\Bigg/\sum_{\tilde{j}\tilde{k}}\left(\frac{(p_{\tilde{j}\tilde{k}})^\mu}{p_{\tilde{j}*}p_{*\tilde{k}}}\right)^{\frac{1}{\mu-1}} \tag{108}$$

When $\mu\to\infty$ for any a $a^{1/(\mu-1)}=1+a/\mu+O(\mu^{-2})$ and (97) follows. For a diagonal P formula (100) is similar to (97) but simpler. ■

APPENDIX B
BUCKETING INFORMATION OF A TENSOR PRODUCT

In this section we prove theorem 4.2:

$$I(P_1\times P_2,\lambda_0,\lambda_1,\mu)=I(P_1,\lambda_0,\lambda_1,\mu)+I(P_2,\lambda_0,\lambda_1,\mu) \tag{109}$$

Proof: Inserting $R=R_1\times R_2$ into definition A.1 proves that $I(P_1\times P_2,\lambda_0,\lambda_1,\mu)\ge I(P_1,\lambda_0,\lambda_1,\mu)+I(P_2,\lambda_0,\lambda_1,\mu)$. Denote $I_1=I(P_1,\lambda_0,\lambda_1,\mu)$, $I_2=I(P_2,\lambda_0,\lambda_1,\mu)$ and $P=P_1\times P_2$:

$$p_{j_1k_1j_2k_2}=p_{1,j_1k_1}p_{2,j_2k_2} \tag{110}$$

For any $\{r_{i,j_1j_2k_1k_2}\}_{i,j_1j_2k_1k_2}$

$$(\mu-1)K(R_{*,\dots\cdot}\|P_{\dots\cdot})+\sum_iK(R_{i,\dots\cdot}\|P_{\dots\cdot})= \\ =(\mu-1)\sum_{j_1k_1}r_{*,j_1k_1**}\ln\frac{r_{*,j_1k_1**}}{p_{1,j_1k_1}}+ \\ +\sum_{i,j_1k_1}r_{i,j_1k_1**}\ln\frac{r_{i,j_1k_1**}}{r_{i,****}p_{1,j_1k_1}}+ \\ +(\mu-1)\sum_{j_1k_1j_2k_2}r_{*,j_1k_1j_2k_2}\ln\frac{r_{*,j_1k_1j_2k_2}}{r_{*,j_1k_1**}p_{2,j_2k_2}}+ \\ +\sum_{i,j_1k_1j_2k_2}r_{i,j_1k_1j_2k_2}\ln\frac{r_{i,j_1k_1j_2k_2}}{r_{i,j_1k_1**}p_{2,j_2k_2}} \tag{111}$$

By definition of I_1 and I_2

$$I_1+(\mu-1)\sum_{j_1k_1}r_{*,j_1k_1**}\ln\frac{r_{*,j_1k_1**}}{p_{1,j_1k_1}}+ \\ +\sum_{i,j_1k_1}r_{i,j_1k_1**}\ln\frac{r_{i,j_1k_1**}}{r_{i,****}p_{1,j_1k_1}}\ge \\ \ge\lambda_0\sum_{i,j_1}r_{i,j_1***}\ln\frac{r_{i,j_1***}}{r_{i,****}p_{1,j_1*}}+ \\ +\lambda_1\sum_{i,k_1}r_{i,*k_1**}\ln\frac{r_{i,*k_1**}}{r_{i,****}p_{1,*k_1}} \tag{112}$$

and

$$r_{*,j_1k_1**}I_2+(\mu-1)\sum_{j_2k_2}r_{*,j_1k_1j_2k_2}\ln\frac{r_{*,j_1k_1j_2k_2}}{r_{*,j_1k_1**}p_{2,j_2k_2}}+ \\ +\sum_{i,j_2k_2}r_{i,j_1k_1j_2k_2}\ln\frac{r_{i,j_1k_1j_2k_2}}{r_{i,j_1,k_1**}p_{2,j_2k_2}}\ge \\ \ge\lambda_0\sum_{i,j_2}r_{i,j_1k_1j_2*}\ln\frac{r_{i,j_1k_1j_2*}}{r_{i,j_1k_1**}p_{2,j_2*}}+ \\ +\lambda_1\sum_{i,k_2}r_{i,j_1k_1*k_2}\ln\frac{r_{i,j_1k_1*k_2}}{r_{i,j_1,k_1**}p_{2,*k_2}} \tag{113}$$

Lemma 7.1 implies

$$I_2+(\mu-1)\sum_{j_1k_1j_2k_2} r_{*,j_1k_1j_2k_2}\ln\frac{r_{*,j_1k_1j_2k_2}}{r_{*,j_1k_1**}p_{2,j_2k_2}}+ \\ +\sum_{i,j_1k_1j_2k_2} r_{i,j_1k_1j_2k_2}\ln\frac{r_{i,j_1k_1j_2k_2}}{r_{i,j_1,k_1**}p_{2,j_2k_2}}\geq \\ \geq\lambda_0\sum_{i,j_1j_2} r_{i,j_1*j_2*}\ln\frac{r_{i,j_1*j_2*}}{r_{i,j_1***}p_{2,j_2*}}+ \\ +\lambda_1\sum_{i,k_1k_2} r_{i,*k_1*k_2}\ln\frac{r_{i,*k_1*k_2}}{r_{i,*,k_1**}p_{2,*k_2}} \quad (114)$$

Together

$$I_1+I_2+(\mu-1)K(R_{*,\cdots}\|P_{\cdots})+\sum_i K(R_{i,\cdots}\|P_{\cdots})\geq \\ \geq\lambda_0\sum_i K(R_{i,\cdot*\cdot*}\|P_{\cdot*\cdot*})+\lambda_1\sum_i K(R_{i,*\cdot*\cdot}\|P_{*\cdot*\cdot}) \quad (115)$$

■

APPENDIX C
PROOF OF THEOREM 5.1

In light of theorem 4.2 it is enough to prove (42) for $d=1$:

$$\ln W\geq\lambda_0\ln n_0+\lambda_1\ln n_1+\mu\ln S-I(P,\lambda_0,\lambda_1,\mu) \quad (116)$$

Proof: Let $(B_{0,0},B_{1,0}),\cdots,(B_{0,T-1},B_{1,T-1})$ be subset pairs. Denote

$$B_i=B_{0,i}\times B_{1,i}\setminus\bigcup_{t=0}^{i-1}B_{0,t}\times B_{1,t} \quad (117)$$

so the success probability is $S=\sum_i p_{B_i}$. Insert

$$r_{i,jk}=\begin{cases}\frac{p_{jk}}{S} & (j,k)\in B_i\\ 0 & \text{otherwise}\end{cases} \quad (118)$$

into definition A.1. Lemma 7.1 implies

$$K(R_{i,\cdot*}\|P_{\cdot*})=\sum_{j\in B_{0,i}} r_{i,j*}\ln\frac{r_{i,j*}}{r_{i,**}p_{j*}}= \\ +r_{i,**}\sum_{j\in B_{0,i}}\frac{r_{i,j*}}{r_{i,**}}\ln\frac{r_{i,j*}}{r_{i,**}p_{j*}}\geq -r_{i,**}\ln p_{B_{0,i}*} \quad (119)$$

Similarly

$$K(R_{i,*\cdot}\|P_{*\cdot})\geq -r_{i,**}\ln p_{*B_{1,i}} \quad (120)$$

Thus

$$\sum_i\left[\lambda_0K(R_{i,\cdot*}\|P_{\cdot*})+\lambda_1K(R_{i,*\cdot}\|P_{*\cdot})\right]\geq \\ \geq-\sum_i r_{i,**}\left(\lambda_0\ln p_{B_{0,i}*}+\lambda_1\ln p_{*B_{1,i}}\right) \quad (121)$$

Recall that the work satisfies $3W\geq W_*=\sum_i W_i$ where

$$W_i=\max\left(n_0p_{B_{0,i}*},\ n_1p_{*B_{1,i}},\ n_0p_{B_{0,i}*}n_1p_{*B_{1,i}}\right) \quad (122)$$

so for $(\lambda_0,\lambda_1)=(0,1),(1,0),(1,1)$

$$\ln W_i\geq\lambda_0\ln(n_0p_{B_{0,i}*})+\lambda_1\ln(n_1p_{*B_{1,i}}) \quad (123)$$

and by linearity it holds for all $(\lambda_0,\lambda_1)\in\mathrm{Conv}(\{(1,0),(0,1),(1,1)\})$. Hence

$$\ln W_i\geq\lambda_0\ln(n_0p_{B_{0,i}*})+\lambda_1\ln(n_1p_{*B_{1,i}}) \quad (124)$$

and thus

$$-\left(\lambda_0\ln p_{B_{0,i}*}+\lambda_1\ln p_{*B_{1,i}}\right)\geq\lambda_0\ln n_0+\lambda_1\ln n_1-\ln W_i \quad (125)$$

Clearly

$$K(R_{*,\cdot\cdot}\|P_{\cdot\cdot})=-\ln S \quad (126)$$

$$\sum_i K(R_{i,\cdot\cdot}\|P_{\cdot\cdot})=-\sum_{ijk} r_{i,jk}\ln(r_{i,**}S)= \\ =-\ln S-\sum_i r_{i,**}\ln r_{i,**} \quad (127)$$

Now definition A.1,(121),(125),(126) and (127) come together:

$$I(P,\lambda_0,\lambda_1,\mu)\geq\lambda_0\sum_i K(R_{i,\cdot*}\|P_{\cdot*})+ \\ +\lambda_1\sum_i K(R_{i,*\cdot}\|P_{*\cdot})-(\mu-1)K(R_{*,\cdot\cdot}\|P_{\cdot\cdot})- \\ -\sum_i K(R_{i,\cdot\cdot}\|P_{\cdot\cdot})\geq \\ \geq\sum_i r_{i,**}\left(\lambda_0\ln n_0+\lambda_1\ln n_1-\ln W_i\right)+ \\ +\mu\ln S+\sum_i r_{i,**}\ln r_{i,**}= \\ =\lambda_0\ln n_0+\lambda_1\ln n_1+\mu\ln S+\sum_i r_{i,**}\ln\frac{r_{i,**}}{W_i} \quad (128)$$

Another call of duty for lemma 7.1 produces

$$\sum_i r_{i,**}\ln\frac{r_{i,**}}{W_i}\geq-\ln W_* \quad (129)$$

Insertion into (128) results in

$$\ln 3W\geq\ln W_*\geq\lambda_0\ln n_0+\lambda_1\ln n_1+\mu\ln S-I(P,\lambda_0,\lambda_1,\mu) \quad (130)$$

In order to get rid of the pesky 3 we concatenate the bucketing code with itself $d\to\infty$ times. Then

$$\ln 3W^d\geq\lambda_0\ln n_0^d+\lambda_1\ln n_1^d+\mu\ln S^d-I(P,\lambda_0,\lambda_1,\mu)d \quad (131)$$

and division by d concludes this proof. ■

APPENDIX D
PROOF OF THEOREM 5.2

Definition D.1: For any probability matrix P let the set $E(P)$ be

$$E(P)=[\cup_{\nu\geq1}E_\nu(P)]^c \quad (132)$$

$$E_\nu(P)=\mathrm{Conv}\left\{\frac{1}{\nu}(\ln n_0,\ln n_1,-\ln\tilde{S},\ln\tilde{W})\right\} \quad (133)$$

where $n_0,n_1>0$ are arbitrary and $0\leq\tilde{S}\leq S,\ \tilde{W}\geq W$. By S,W we denote the success probability and work of some ν dimensional bucketing code for $P_i=P$.

In the previous appendix we have shown that $E(P)\subset D(P)=D(P)+\mathrm{Cone}((0,0,1,-1))$. Theorem D.3 will

prove a strong version of the converse $E(P) \supset D(P) + \mathrm{Cone}((0,0,1,-1))$. It is the heart of this appendix, the rest are technicalities.

When a bucketing code contains only one bucket pair, we do not have to worry about the bucketing work because it is bounded from above by $\max(n_0, n_1)$. Hence it makes sense to define

Definition D.2: For any probability matrix P let the sets $\tilde{E}(P), \tilde{E}_\nu(P)$ be defined by (132),(133) but with $S = p_{B_0B_1}$, $W = n_0 n_1 p_{B_0 *} p_{* B_1}$ for some single bucket pair ($T = 1$) bucketing code.

Definition D.3: For any probability matrix P

$$\rho(P) = -\ln \min_{\substack{0\le j<b_0\\ 0\le k\le b_1\\ p_{jk}>0}} p_{jk} \tag{134}$$

The following result shows that the convex hull in (133) is for convenience only.

Lemma D.1: For any vector $(m_0, m_1, s, w) \in E_\nu(P)$ there exists a ν^2 dimensional bucketing code with vector

$$(m_0, m_1, s, (1 + 8\rho(P)/\nu)w) \tag{135}$$

There also exits a ν^2 dimensional bucketing code with vector

$$(m_0, m_1, s + 4\rho(P)/\nu, w) \tag{136}$$

Similar statements hold for $\tilde{E}_\nu(P)$.

Proof: Let $(m_0, m_1, s, w) \in E_\nu(P)$. By Caratheodory's theorem

$$(m_0, m_1, s, w) = \sum_{i=1}^{5} \alpha_i (m_{i,0}, m_{i,1}, s_i, w_i) \tag{137}$$

where $\alpha_i \ge 0$, $\sum_{i=1}^5 \alpha_i = 1$ and $(m_{i,0}, m_{i,1}, s_i, w_i)$ is the vector of a ν dimensional bucketing code. Without restricting generality we can exclude codes with more than $e^{\nu \max(m_0, m_1, m_0+m_1)}$ work, hence $w_i \le w + 2\rho(P)$. Let us concatenate ν_1 copies of the first code with ν_2 copies of the second code etc, where $\sum_{i=1}^5 \nu_i = \nu$ and $\nu_i \ge \lfloor \alpha_i \nu \rfloor$. The extra $\sum_{i=1}^5 (\alpha_i \nu - \lfloor \alpha_i \nu \rfloor) \le 4$ copies go to the lowest s_i code, so $-\frac{1}{\nu^2} \ln S \le s$ and $\frac{1}{\nu^2} \ln W \le w + \frac{8\rho(P)}{\nu} w$. Thus we attained (135).

In order to get (136) we exclude codes with zero success probability, hence $s_i \ge \rho(P)$, and give the extra weight to the lowest w_i code. ∎

We will use the following multinomial estimate.

Lemma D.2: For any probability distribution $\{r_i \ge 0\}_{0\le i<I}$, $r_* = 1$ and and an integer $\nu \ge 1$ there exist nonnegative integers $\{\nu_i \ge 0\}_{0\le i<I}$ such that $\nu_* = \nu$ and

$$r_i \nu - 1 < \nu_i < r_i \nu + 1 \tag{138}$$

Moreover these inequalities imply

$$\nu^{1-2I} \le \frac{\nu_*!}{\prod_{i=0}^{I-1} \nu_i!} \prod_{i=0}^{I-1} r_i^{r_i \nu} \le \nu^I \tag{139}$$

Proof: Let $\nu_i = \lfloor r_i \nu \rfloor + \epsilon_i$ where $\epsilon_i = 0, 1$ in general, $\epsilon_i = 0$ when $r_i \nu$ is integral and $\epsilon_* = \sum_i (r_i \nu - \lfloor r_i \nu \rfloor)$. Thus we get (138). The case of $\nu = 1$ is trivial so let $\nu \ge 2$. Estimate (139) follows from the well known multinomial inequality

$$\nu^{1-I} \le \frac{\nu_*!}{\prod_{i=0}^{I-1} \nu_i!} \prod_{i=0}^{I-1} \left(\frac{\nu_i}{\nu}\right)^{\nu_i} \le 1 \tag{140}$$

and the easy bound

$$|x \ln x - y \ln y| \le -|x-y| \ln |x-y| \tag{141}$$

valid for all any $0 \le x, y \le 1$, $|x - y| \le 1/2$. ∎

Theorem D.3: For any probability matrix P and $\nu \ge 1$

$$D(P) + b_0 b_1 \frac{\rho(P) + \ln \nu}{\nu} (0,0,2,3) \subset \\ \subset E_\nu(P) \cup \Big[(E_\nu(P) \cap \tilde{E}_\nu(P)) + \mathrm{Cone}((0,0,1,-1)) \Big] \tag{142}$$

Proof: The single big bags pair code

$$B_0 = \{0, 1, \ldots, b_0 - 1\},\ B_1 = \{0, 1, \ldots, b_1 - 1\} \tag{143}$$

shows that

$$\mathrm{ConvCone}(\{(1,0,0,1), (0,1,0,1), (-1,-1,0,-1), \\ , (0,0,1,0), (0,0,0,1)\}) \subset E_1(P) \cap \tilde{E}_1(P) \tag{144}$$

Let $\{r_{i,jk}\}_{ijk}$ satisfy $r_{i,jk} \ge 0$, $r_{*,**} = 1$. We will prove that

$$\left(\sum_i K(R_{i,\cdot *} \| P_{\cdot *}),\ \sum_i K(R_{i,* \cdot} \| P_{* \cdot}), \right. \\ \left. , \sum_i K(R_{i,\cdot\cdot} \| P_{\cdot\cdot}),\ 0 \right) + \\ + b_0 b_1 \frac{\rho(P) + \ln \nu}{\nu} (0,0,2,2) \in E_\nu(P) \cap \tilde{E}_\nu(P) \tag{145}$$

and

$$\left(\sum_i K(R_{i,\cdot *} \| P_{\cdot *}), \sum_i K(R_{i,* \cdot} \| P_{* \cdot}), K(R_{*,\cdot\cdot} \| P_{\cdot\cdot}), \right. \\ \left. , -K(R_{*,\cdot\cdot} \| P_{\cdot\cdot}) + \sum_i K(R_{i,\cdot\cdot} \| P_{\cdot\cdot}) \right) + \\ + b_0 b_1 \frac{\rho(P) + \ln \nu}{\nu} (0,0,2,3) \in E_\nu(P) \tag{146}$$

Formulas (144),(146) together with lemma A.1 prove

$$D(P) + 3 b_0 b_1 \frac{\ln \nu}{\nu} (0,0,1,1) \subset E_\nu(P) + \mathrm{Cone}((0,0,1,-1)) \tag{147}$$

Formula (142) is proven as follows. For any $(m_0, m_1, s, w) \in E(P) + \mathrm{Cone}((0,0,1,-1))$ let α be the minimal $\alpha \ge 0$ such that $(m_0, m_1, s - \alpha, w + \alpha) \in E(P)$. When $\alpha = 0$ we are done. Otherwise $(m_0, m_1, s - \alpha, w + \alpha)$ is in the closure of the convex hull of the $(\sum_i K(R_{i,\cdot *} \| P_{\cdot *}),\ \sum_i K(R_{i,* \cdot} \| P_{* \cdot}),\ \sum_i K(R_{i,\cdot\cdot} \| P_{\cdot\cdot}),\ 0)$ points, so we get (142).

It remains to prove (145),(146). Lemma D.2 provides us with nonnegative $\{\nu_{i,jk}\}_{i,jk}$. Let us define a bucket pair

$$B_{0,0} = \left\{ x_0 \ \middle|\ \forall ij \sum_{l=c_i+1}^{c_{i+1}} (x_{0,l} == j) \ = \ \nu_{i,j*} \right\} \tag{148}$$

$$B_{0,1} = \left\{ x_1 \;\middle|\; \forall ik \sum_{l=c_i+1}^{c_{i+1}} (x_{1,l} == k) \;=\; \nu_{i,*k} \right\} \quad (149)$$

where $c_i = \sum_{l=0}^{i-1} \nu_{i,**}$ In words we want x_0 to contain exactly $\nu_{0,j*}$ j-values in its first $\nu_{0,**}$ coordinates, etc. The bucket probabilities are

$$p_{B_{0,0*}} = \prod_i \left[\frac{\nu_{i,**}!}{\prod_j \nu_{i,j*}!} \prod_j p_{j*}^{\nu_{i,j*}} \right] \quad (150)$$

$$p_{*B_{0,1}} = \prod_i \left[\frac{\nu_{i,**}!}{\prod_k \nu_{i,*k}!} \prod_k p_{*k}^{\nu_{i,*k}} \right] \quad (151)$$

Let us add $T-1$ similar buckets. They are generated by randomly permuting the coordinates $1, 2, \ldots, \nu$. A lower bound of the average success probability of this random bucketing code is

$$\mathrm{E}[S] \geq U \left[1 - (1 - V/U)^T \right] \geq U e^{-TV/U} \quad (152)$$

where

$$U = \frac{\nu!}{\prod_{jk} \nu_{*,jk}!} \prod_{jk} p_{jk}^{\nu_{*,jk}} \quad (153)$$

is the probability that a matched pair obtains coordinate pair (j,k) exactly $\nu_{*,jk}$ times, and

$$V = \prod_i \left[\frac{\nu_{i,**}!}{\prod_{jk} \nu_{i,jk}!} \prod_{jk} p_{jk}^{\nu_{i,jk}} \right] \quad (154)$$

is the probability that a matched pair obtains coordinate pair (j,k) exactly $\nu_{i,jk}$ times in coordinate subset number i. Of course there exists a deterministic code at least as successful as the average code.

Lemma D.2 implies

$$-c \leq -\frac{1}{\nu} \ln p_{B_{0,0*}} - \sum_i K(R_{i,\cdot *} \| P_{*}) \leq 2c - \frac{\rho(P)}{\nu} \quad (155)$$

$$-c \leq -\frac{1}{\nu} \ln p_{*B_{0,1}} - \sum_i K(R_{i,*\cdot} \| P_{*\cdot}) \leq 2c - \frac{\rho(P)}{\nu} \quad (156)$$

$$-c \leq \frac{-1}{\nu} \ln U - K(R_{*,\cdot\cdot} \| P_{\cdot\cdot}) \leq 2c - \frac{\rho(P)}{\nu} \quad (157)$$

$$-c \leq \frac{-1}{\nu} \ln V - \sum_i K(R_{i,\cdot\cdot} \| P_{\cdot\cdot}) \leq 2c - \frac{\rho(P)}{\nu} \quad (158)$$

where $c = b_0 b_1 \frac{\rho(P) + \ln \nu}{\nu}$. Let

$$\ln n_0 = \sum_i K(R_{i,\cdot *} \| P_{*}) \quad (159)$$

$$\ln n_1 = \sum_i K(R_{i,*\cdot} \| P_{*\cdot}) \quad (160)$$

In order to establish (145) we choose $T = 1$. In order to establish (146) we choose $T = \lceil U/V \rceil$. ■

Lemma D.4:

$$\oplus_{i=1}^d D(P_i) + (0,0,o(1),o(1)) \subset \oplus_{i=1}^d E_\nu(P_i) \cup \\ \cup \left[\oplus_{i=1}^d \tilde{E}_\nu(P_i) + \mathrm{Cone}((0,0,1,-1)) \right] \quad (161)$$

where $\oplus_{i=1}^d D(P_i) = D(P_1) + \cdots + D(P_d)$ etc., and $o(1) \to 0$ as $\nu \to \infty$.

Proof: Let $(m_0, m_1, s, w) \in \oplus_{i=1}^d D(P_i)$. We know that there exist $\alpha \geq 0$, $(m_{i,0}, m_{i,1}, s_i, w_i) \in E_\nu(P_i)$ such that

$$(m_0, m_1, s - \alpha, w + \alpha) + (0,0,o(1),o(1)) = \\ = \sum_{i=1}^d (m_{i,0}, m_{i,1}, s_i, w_i) \quad (162)$$

Let $\alpha_i \geq 0$ be the maximal values such that $(m_{i,0}, m_{i,1}, s_i + \alpha_i, w_i - \alpha_i) \in E_\nu(P_i)$. Theorem D.3 implies that

$$(m_{i,0}, m_{i,1}, s_i + \alpha_i, w_i - \alpha_i) + (0,0,o(1),o(1)) \in \\ \in E_{nu}(P_i) \cap \tilde{E}_\nu(P_i) \quad (163)$$

Denote $\beta = -\alpha + \sum_{i=1}^d \alpha_i$. When $\beta \leq 0$

$$(m_0, m_1, s, w) + (0,0,o(1),o(1)) = \\ = -\beta(0,0,1,-1) + \sum_{i=1}^d (m_{i,0}, m_{i,1}, s_i + \alpha_i, w_i - \alpha_i) \quad (164)$$

so $(m_0, m_1, s, w) + (0,0,o(1),o(1)) \in \oplus_{i=1}^d \tilde{E}_\nu(P_i) + \mathrm{Cone}((0,0,1,-1)$. When $\beta \geq 0$

$$(m_0, m_1, s, w) + (0,0,o(1),o(1)) = \\ = \sum_{i=1}^d \left[\frac{\beta}{\alpha + \beta} (m_{i,0}, m_{i,1}, s_i, w_i) + \right. \\ \left. + \frac{\alpha}{\alpha + \beta} (m_{i,0}, m_{i,1}, s_i + \alpha_i, w_i - \alpha_i) \right] \quad (165)$$

so $(m_0, m_1, s, w) + (0,0,o(1),o(1)) \in \oplus_{i=1}^d E_\nu(P_i)$. ■

Definition D.4: For any probability matrix P let $\tilde{H}(P)$ be the half space

$$\tilde{H}(P) = \{ (m_0, m_1, s, w) \mid w \geq m_0 + m_1 - \omega(P) \} \quad (166)$$

Lemma D.5:

$$\left[\tilde{E}_\nu(P) + \mathrm{Cone}((0,0,1,-1)) \right] \cap \tilde{H}(P) \subset \tilde{E}_\nu(P) \quad (167)$$

Proof: Let (m_0, m_1, s, w) be in the left side set. For some $\alpha \geq 0$

$$(m_0, m_1, s - \alpha, w + \alpha) \in \tilde{E}_\nu(P) \quad (168)$$

Denote

$$\beta = w - m_0 - m_1 + \omega(P) \geq 0 \quad (169)$$

Clearly $p_{B_0 B_1} \leq \frac{p_{j_1 k_1}}{p_{j_1 * p_{* k_1}}} p_{B_0 *} p_{* B_1}$ so $-s \leq \ln \frac{p_{j_1 k_1}}{p_{j_1 * p_{* k_1}}} + w - m_0 - m_1$ which can be written as

$$\beta + s + \ln p_{j_1 k_1} \geq 0 \quad (170)$$

The single bucket pair code $B_0 = \{(j_1)\}, B_1 = \{(k_1)\}$ shows that

$$(m_0, m_1, -\ln p_{j_1 k_1}, m_0 + m_1 + \ln p_{j_1 * p_{* k_1}}) = \\ = (m_0, m_1, -\ln p_{j_1 k_1}, w - \beta) \in \tilde{E}_1(P) \quad (171)$$

Hence

$$(m_0,m_1,s,w)=\frac{\beta}{\alpha+\beta}(m_0,m_1,s-\alpha,w+\alpha)+ \\ +\frac{\alpha}{\alpha+\beta}(m_0,m_1,-\ln p_{j_1k_1},w-\beta)+ \\ +\frac{\alpha}{\alpha+\beta}(0,0,\beta+s+\ln p_{j_1k_1},0)\in\tilde{E}_\nu(P) \quad (172)$$

■

Lemma D.6:

$$\left[\oplus_{i=1}^d\tilde{E}_\nu(P_i)+\mathrm{Cone}((0,0,1,-1))\right]\cap \\ \cap\oplus_{i=1}^d\tilde{H}(P_i)\subset\oplus_{i=1}^d\tilde{E}_\nu(P_i) \quad (173)$$

Proof: Let (m_0,m_1,s,w) belong to left side set. There exist $\alpha\geq 0$, $(m_{i,0},m_{i,1},s_i,w_i)\in\tilde{E}_\nu(P_i)$ such that

$$(m_0,m_1,s-\alpha,w+\alpha)=\sum_{i=1}^d(m_{i,0},m_{i,1},s_i,w_i) \quad (174)$$

$$w\geq m_0+m_1-\sum_{i=1}^d\omega(P_i) \quad (175)$$

Let $\alpha_i\geq 0$ be the maximal value such that $(m_{i,0},m_{i,1},s_i+\alpha_i,w_i-\alpha_i)\in\tilde{E}_\nu(P_i)$. Lemma D.5 implies that $w_i-\alpha_i\leq m_{i,0}+m_{i,1}-\omega(P_i)$ which is rearranged as

$$\alpha_i\geq w_i-m_{i,0}-m_{i,1}+\omega(P_i) \quad (176)$$

and summed up to give

$$\beta=-\alpha+\sum_{i=1}^d\alpha_i\geq w-m_0-m_1+\sum_{i=1}^d\omega(P_i)\geq 0 \quad (177)$$

Hence

$$(m_0,m_1,s,w)=\sum_{i=1}^d\left[\frac{\beta}{\alpha+\beta}(m_{i,0},m_{i,1},s_i,w_i)+ \right. \\ \left. +\frac{\alpha}{\alpha+\beta}(m_{i,0},m_{i,1},s_i+\alpha_i,w_i-\alpha_i)\right] \quad (178)$$

■

Lemmas D.4 and D.6 combine into

Theorem D.7:

$$\left(\oplus_{i=1}^d D(P_i)\cap\oplus_{i=1}^d\tilde{H}(P_i)\right)+(0,0,o(1),o(1))\subset \\ \subset\oplus_{i=1}^d E_\nu(P_i)\cup\oplus_{i=1}^d\tilde{E}_\nu(P_i) \quad (179)$$

Proof of theorem 5.2.

Proof: Let $\nu>1$ be a "large enough" integer (to be determined). We are given $(m_0,m_1,s,w)\in\oplus_{i=1}^d D(P_i)\cap\oplus_{i=1}^d\tilde{H}(P_i)$, $s\geq 0$ and $w\geq m_0,m_1\geq 0$. Theorem D.7 tells that either $(m_0,m_1,s,w)+(0,0,o(1),o(1))\in\oplus_{i=1}^d E_\nu(P_i)$ or a similar statement holds for $\tilde{E}_\nu(P_i)$.

First consider the homogeneous $P_i=P$ case. Lemma D.1 implies the existence of a ν^2 dimensional bucketing code with vector $(m_0,m_1,s+o(1),w)$. The $\lfloor d/\nu^2\rfloor$'th tensor power of that code attains all theorem 5.2's claims. The n_0+n_1 vector pointers are the depth first search pointers, as discussed in section II.

When the P_i's are not equal, we replace them with approximations taken from a prearranged constant set of matrices. For $\delta>0$ $0\leq j<b$ let $P(\delta)$ be defined by $p_{jk}(\delta)=c\mathrm{e}^{-\lceil-(\ln p_{jk})/\delta\rceil\delta}$ where the normalizing c enforces $p_{**}(\delta)=1$. The work and success are changed by at most $\mathrm{e}^{-\delta d}\leq W(\delta)/W,S(\delta)/S\leq\mathrm{e}^{\delta d}$. In particular $(m_0,m_1,s+2\delta d,w+2\delta d)\in\oplus_{i=1}^d D(P_i(\delta))\cap\oplus_{i=1}^d\tilde{H}(P_i(\delta))$. Requiring $p_{i,jk}=0$ or $p_{i,jk}>\epsilon$ for all $1\leq i\leq d$, $0\leq j<b_0$, $0\leq k<b_1$ bounds the number of possible $P(\delta)$ by $(-(\ln\epsilon)/\delta+1)^{b_0b_1}$. For each $P(\delta)$ construct a bucketing code. Then replace $P_1(\delta),\ldots,P_d(\delta)$ by $P_1,\ldots,P_d$, decreasing success and increasing work again by a factor. Taking small enough δ completes this proof. ■

ACKNOWLEDGMENT

This paper is dedicated to my wife Edith whose support made it possible, and to Benjamin Weiss who taught me what mathematical information means. I also thank Uri Zwick and the referees for suggesting clarifications.

REFERENCES

[1] N. Alon Private Communication.
[2] A. Andoni, P. Indyk *Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions* FOCS 2006.
[3] A. Broder. *Identifying and Filtering Near-Duplicate Documents* Proc. FUN, 1998.
[4] M. Datar, P. Indyk, N. Immorlica and V. Mirrokni *Locality-Sensitive Hashing Scheme Based on p-Stable Distributions* Proc. Sympos. on Computational Geometry, 2004.
[5] M. Dubiner *A Heterogeneous High Dimensional Approximate Nearest Neighbor Algorithm* To be Published.
[6] M. Dubiner *Computing Bucketing Information* To be Published.
[7] C.Gennaro, P.Savino and P.Zezula *Similarity Search in Metric Databases through Hashing* Proc. ACM workshop on multimedia, 2001.
[8] P. Indyk and R. Motwani. *Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality* STOC 1998.
[9] R.M. Karp, O. Waarts, and G. Zweig. *The Bit Vector Intersection Problem* FOCS 1995.
[10] U. Manber *Finding similar files in a large file system* Proc. of the USENIX Winter 1994 Technical Conference
[11] R. Motwani, A. Naor and R. Panigrahy *Lower Bounds on Locality Sensitive Hashing* SCG'06
[12] R. Paturi, S. Rajasekaran and J. Reif *The Light Bulb Problem* Information and Computation, Vol 117, Issue 2, 1995.

Moshe Dubiner Moshe Dubiner has been in the Department of Applied Mathematics at the Tel Aviv University, where he worked on approximation theory. He has been a quantitative analyst, and is now a researcher at Google.