Department of Computer Science Technical Reports

Department of Computer Science

1981

Some Results on Perfect, Static-Key Hashing

Eric Dittert

Michael J. O'Donnell

Report Number:
81-390

Some Results on Perfect, Static-Key Hashing[†]

Eric Dittert

Michael J. O'Donnell

Department of Computer Sciences
Purdue University
West Lafayette, Indiana

CSD-TR-390

Introduction

Often a computer program needs to accept as all or part of its input a sequence of character strings and decide, for each string, whether that string is a member of some finite set of known strings. The set of known strings may be nonempty when the program starts and may change as the program receives input. The strings, both known and otherwise, are generally referred to as **keys**. Testing a key for membership in the set of known keys is called a **search**, adding a key to the set of known keys is called an **insertion**, and removing a key from the set is a **deletion**.

Many different schemes have been developed to handle this computational task. These include linear search of an unordered table, binary search of an ordered table, B-trees, tries, various forms of string pattern matching, and hashing. The choice of one scheme over another for a certain application generally depends on the size of the set of known keys, and the relative numbers and order of searches, insertions, and deletions, since each scheme is efficient in some situations, and inefficient or inapplicable in others.

Hashing refers to schemes which perform an insertion by computing some (simple) arithmetic function of the key to be inserted and using the result as the location in the table of known keys at which the key should be stored. The view of the hashing procedure as dropping keys into various locations in the table has led to the locations in the table (whether containing a key or not) being referred to as **buckets**. A search for a key is performed by computing the same function on the key to be searched for, and then comparing the key with whatever, if anything, is currently stored in the indicated bucket. A **collision** occurs when two keys to be inserted are mapped by the function to the same bucket. Each hashing scheme must include some method for dealing with collisions. There are many interesting possible methods, but we will be concerned with a special case in which this is not an issue.

[†] This work supported by NSF grant MCS 78-01812

Most of the research on hashing has dealt with the general case in which search operations and insertion operations may be intermingled in an arbitrary manner. However, if the insertion operations are guaranteed to precede all the search operations then all the keys which will ever be in the table are known in advance. In this case one might try to find a function mapping these keys, without collisions, into a table not much larger than the set of keys.

Perfect hashing

The following definitions are due to Sprugnoli [Spr 77].

Definition 1: A hashing function is a **perfect hashing function** (or **phf**) for a set of keys iff the function is 1-1 on that set of keys, *i.e.*, there are no collisions on those keys.

Definition 2: A hashing function is a **minimal phf** for a set of keys iff the function maps the keys in a 1-1 fashion onto the buckets $0,1,...,k-1$, where k is the number of keys in the set.

Thus if one has a minimal phf for a set of keys, the hash table need have only as many entries as there are keys in the set. A **nearly minimal phf** is a function which uses a table not much larger than the set of keys, where "not much larger than" might be interpreted, for example, as "not larger than a constant factor times".

Every set of keys has, of course, several minimal phfs; the problem is that we may not know how to compute any of them in constant time, as we would like for a hashing function. Several articles ([Spr 77], [Cic 80]) recount procedures which take as input a set of keys and try to produce a minimal or nearly minimal phf for that set of keys in a form which is known *a priori* to be computable in constant time. The success of these procedures varies with the cleverness of the authors, but none of them is guaranteed to work for every set of keys. Hence, the nature and complexity of general methods for producing minimal or nearly minimal phfs are still open.

Alternate approaches to the general problem include that of Tarjan and Yao [Tar 79] which produces minimal phfs computable in $O(log_k P)$ time, where P is the size of the universal set from which k keys are drawn. Also, Carter and Wegman [Car 77] study classes of functions with the property that given a set of keys, S, a function chosen at random from the class will be, on the average, a good hashing function for S.

In this report we address the question of what restrictions on the class of functions considered would guarantee that the class contained at least one phf for each set of keys (disregarding, for the moment, how to compute the functions in some way faster than table look-up).

Formalizing some questions

We shall take the set of possible keys and the set of buckets each to be an initial segment of the natural numbers. In addition, we specify the size of the sets of keys for which phfs are to found. We shall use the following symbols:

P	the number of possible keys, those being $\{0,1,\dots,P-1\}$
b	the number of buckets, those being $\{0,1,\dots,b-1\}$
k	the number of keys in a key set
$b^P =$	$\{f \mid f:\{0,1,\dots,P-1\} \rightarrow \{0,1,\dots,b-1\}\}$

Also, **R** will denote the real numbers, $\mathbf{R}^+$ the nonnegative real numbers, and **N** the natural numbers. And we shall need the following definitions.

Definition 3: A **(k,P)-set** is a k-element subset of $\{0,1,\dots,P-1\}$.

Definition 4: A function, $f \in b^P$ **covers** a (k,P)-set, S, iff f is a phf for S. Also, $G \subseteq b^P$ **covers** a (k,P)-set, S, iff for at least one $g \in G$, g covers S.

We can now state precisely the questions to which we wish to obtain answers.

> Question 1: Given P, b, k and an integer n, what choice of n functions, $f_1,f_2,\dots,f_n \in b^P$ will maximize the number of (k,P)-sets covered by $\{f_1,\dots,f_n\}$, and how many (k,P)-sets will it cover?
>
> Question 2: Given P, b,and k, what is the smallest integer n such that every (k,P)-set is covered by some $\{f_1,\dots,f_n\} \subseteq b^P$?

The function *cov*

Definition 5: $cover_k(\{f_1,\dots,f_n\})$ is the number of (k,P)-sets covered by $\{f_1,\dots,f_n\}$.

Our first aim is to obtain a convenient expression for $cover_k(\{f_1,\dots,f_n\})$ in terms of quantities easily determinable from the functions themselves.

If we have just one function, $f \in b^P$, then we can count the number of (k,P)-sets covered by $\{f\}$ in the following way. First, let "bucket-set i" denote $f^{-1}(i)$, the set of keys mapped by f to i. Now select any k *distinct* buckets $i_1,\dots,i_k$, and the corresponding bucket-sets. Then if we form a (k,P)-set by picking one key from each of the selected bucket-sets, that (k,P)-set is covered by f. For a particular collection of k buckets, $I=\{i_1,\dots,i_k\}$, the number of (k,P)-sets that we can form in this way is just the product of the cardinalities of the corresponding bucket-sets:

$$\prod_{j=1}^{k} |\text{ bucket-set } i_j \,| = \prod_{i\in I} | f^{-1}(i) \,|.$$

If we then sum over all ways of picking k buckets, we will have counted exactly once each (k,P)-set that f covers. So

$$cover_k(\{f\}) = \sum_{\substack{I\subseteq\{0,1,\ldots,b-1\}\\ |I|=k}} \prod_{i\in I} |f^{-1}(i)|.$$

For sets $F=\{f_1,\ldots,f_n\}\subseteq b^P$ containing more than one function the situation is a little more complicated. It will not do to just sum the preceding formula over all the functions in the set, since a (k,P)-set could be covered by more than one function, but should only be counted once. One way to overcome this difficulty is to generalize our notion of bucket-sets (the bucket-sets of single functions being hereafter referred to as simple bucket-sets). Each generalized bucket-set is the intersection of a collection of simple bucket-sets, one from each function. We name a generalized bucket-set by a vector, $\mu=<\mu(1),\ldots,\mu(n)>$, which indicates from which simple bucket-sets it was derived. More precisely, the generalized bucket-set Z_μ is given by

$$\begin{aligned} Z_\mu(F) = Z_{\mu(1),\ldots,\mu(n)}(F) &= \bigcap_{j=1}^{n} \text{bucket-set}\,\mu(j) \text{ from function } j \\ &= \bigcap_{j=1}^{n} f_j^{-1}(\mu(j)). \end{aligned}$$

Unfortunately, it is not the case that picking one key from each of k distinct generalized bucket-sets will guarantee that the (k,P)-set so formed is covered by F. The essence of the complication is that for F to cover a (k,P)-set, some *single* function in F must cover that (k,P)-set. If we select k generalized bucket-sets, $Z_{\mu_1},\ldots,Z_{\mu_k}$, then the (k,P)-sets formed by picking one key from each will be covered by F just if there is some $f\in F$ such that $Z_{\mu_1},\ldots,Z_{\mu_k}$ are portions of k distinct simple bucket-sets of f. One can visualize this condition in terms of the names of the generalized bucket-sets, $\mu_1,\ldots,\mu_k$, as follows. Align the vectors $\mu_1,\ldots,\mu_k$ one beneath the other:

$$\begin{array}{cccc} < \mu_1(1) & \cdots & \mu_1(n) & > \\ < \mu_2(1) & \cdots & \mu_2(n) & > \\ \vdots & & \vdots & \\ < \mu_k(1) & \cdots & \mu_k(n) & > \end{array}$$

Then F covers the (k,P)-sets formed by picking one key from each of $Z_{\mu_1},\ldots,Z_{\mu_k}$ if and only if some column of the array above comprises k distinct entries.

So to compute $cover_k(\{f_1,\ldots,f_n\})$ we take, as in the case of one function, products of cardinalities in collections of k (generalized) bucket-sets. But we sum only over the collections of bucket-sets having the property described above. To make writing this easier, we introduce

$$diff_k(\mu_1,\ldots,\mu_k) = |\{j : \mu_1(j),\ldots,\mu_k(j) \text{ are pairwise distinct}\}|$$

Then we have

$$cover_k(\{f_1,\ldots,f_n\}) = \frac{1}{k!} \sum_{diff_k(\mu_1,\ldots,\mu_k)\geq 1} \prod_{j=1}^{k} |Z_{\mu_j}(\{f_1,\ldots,f_n\})|.$$

The factor $\frac{1}{k!}$ reflects the fact that each collection of k bucket-sets, $Z_{\xi_1},\ldots,Z_{\xi_k}$, can be ordered in $k!$ different ways to appear as $Z_{\mu_1},\ldots,Z_{\mu_k}$ in the formula.

Now, since only the cardinalities of the various Z_μ are important let us define

$$a_\mu(\{f_1,\ldots,f_n\}) = |Z_\mu(\{f_1,\ldots,f_n\})| = |\bigcap_{j=1}^{n} f_j^{-1}(\mu(j))|.$$

Then $A(\{f_1,\ldots,f_n\})=[a_\mu(\{f_1,\ldots,f_n\})]$ will be an n-dimensional $(b \times b \times \cdots \times b)$ array of nonnegative integers with $\sum_\mu a_\mu = P$ (since each key is counted in exactly one a_μ). We will refer to A as the **key distribution** of $\{f_1,\ldots,f_n\}$.

We can give one result immediately in regard to Question 2:

Proposition 1: If $\{f_1,\ldots,f_n\} \subseteq b^P$ covers all (k,P)-sets (for some $2 \leq k \leq b$), then $A(\{f_1,\ldots,f_n\})$ is such that $\forall \mu$, $a_\mu(\{f_1,\ldots,f_n\}) \in \{0,1\}$.

> Proof: Suppose for some μ, $a_\mu(\{f_1,\ldots,f_n\}) \geq 2$. Then there are two keys, s_1 and s_2, such that $\forall f \in \{f_1,\ldots,f_n\}$, $f(s_1)=f(s_2)$. Pick any (k,P)-set, S, containing both s_1 and s_2. Since no $f \in \{f_1,\ldots,f_n\}$ is 1-1 on S, $\{f_1,\ldots,f_n\}$ does not cover S.

Corollary: If $\{f_1,\ldots,f_n\} \subseteq b^P$ covers all (k,P)-sets (for some $2 \leq k \leq b$), then $n \geq \lceil \log_b P \rceil$.

> Proof: By definition there are b^n elements in $A(\{f_1,\ldots,f_n\})$ and their sum is P. Since $\{f_1,\ldots,f_n\}$ covers all (k,P)-sets, each element is either 0 or 1. Hence
>
> $$P = \sum_\mu a_\mu \leq b^n \cdot 1 = b^n$$
>
> and so $n \geq \log_b P$. Since n is an integer, $n \geq \lceil \log_b P \rceil$.

Definition 6a: For $n \geq 1, b \geq 1$, and $P>0$, let $\Omega_{n,b,P}$ denote the space of all n-dimensional, $b \times b \times \cdots \times b$ arrays, A, with entries which are real numbers such that $\sum_\mu a_\mu = P$.

Definition 6b: For $n \geq 1, b \geq 1$, and $P>0$, let $\Omega^+_{n,b,P}$ denote the space of all n-dimensional, $b \times b \times \cdots \times b$ arrays, A, with entries which are nonnegative real numbers such that $\sum_\mu a_\mu = P$.

Definition 6c: For $n \geq 1, b \geq 1$, and $P > 0$, let $\Omega^{\mathbf{N}}_{n,b,P}$ denote the space of all n-dimensional, $b \times b \times \cdots \times b$ arrays, A, with entries which are nonnegative integers such that $\sum_{\mu} a_{\mu} = P$.

Observing that

1) every $A \in \Omega^{\mathbf{N}}_{n,b,P}$ is the key distribution of some set of n functions in b^P; and
2) from a key distribution we can easily find a set of functions having that key distribution,

we will cease to mention functions explicitly and define

$$cov_{n,b,k}(A) = \frac{1}{k!} \sum_{diff_k(\mu_1,\ldots,\mu_k) \geq 1} \prod_{j=1}^{k} a_{\mu_j}.$$

Then we can answer Question 1 if we can find the maxima of $cov_{n,b,k}(A)$ for $A \in \Omega^{\mathbf{N}}_{n,b,P}$.

Unfortunately, we have been able to do this only for the case of one function ($n=1$), and for the case of two keys per set ($k=2$). In light of this difficulty we have tried to obtain at least an approximate answer for $n>1$ by extending the domain of $cov_{n,b,k}(A)$ to arrays of real numbers, (*i.e.*, to $\Omega_{n,b,P}$) (which involves no work) and seeking a global maximum for $A \in \Omega_{n,b,P}$. At this, we have been successful for the case $k=2$. However, the example at the end of the next section shows that $cov_{n,b,k}(A)$ may be unbounded if A contains negative values and k is greater than 2. Hence, we will ultimately have to restrict our attention to $\Omega^{+}_{n,b,P}$, though this should not sadden us greatly in view of our intended application.

Results for the case of one function ($n=1$)

If we have just one function, then the key distribution is a vector, $A=[a_1, \ldots, a_b]$, and

$$cov_{1,b,k}(A) = \sum_{\substack{I \subseteq \{1,\ldots,b\} \\ |I|=k}} \prod_{i \in I} a_i$$

Definition 7: For each real constant c let $[c]_{\mathbf{n,b}}$ denote the n-dimensional $b \times b \times \cdots \times b$ array, all the entries of which have the value c.

Definition 8: $A \sim_{\mathbf{N}} [c]_{n,b}$ (read A approximates $[c]_{n,b}$ with integers) if and only if the sum of the elements of A is $c \cdot b^n$ and each element of A is equal to either $\lfloor c \rfloor$ or $\lceil c \rceil$.

Definition 9: A function $f \in b^P$ is an **even-sprinkling function** iff its key distribution, A, is such $A \sim_{\mathbf{N}} \left[\frac{P}{b}\right]_{1,b}$. Thus the key distribution of an even-sprinkling function is in some sense the best integer approximation to $\left[\frac{P}{b}\right]_{1,b}$.

Anderson and Sprugnoli [And 79] considered the special case for one function when $k=b$. They showed that $cov_{1,b,b}(A)$, $A \in \Omega^{+}_{1,b,P}$, has a global maximum at $A=\left[\frac{p}{b}\right]_{1,b}$; and that $cov_{1,b,b}(A)$, $A \in \Omega^{\mathbf{N}}_{1,b,P}$, is maximized whenever $A \sim_{\mathbf{N}} \left[\frac{p}{b}\right]_{1,b}$. Berman, Bock, and Plank [Ber 81] proved that, in fact, for any $k \le b$, $A=\left[\frac{p}{b}\right]_{1,b}$ maximizes $cov_{1,b,k}(A)$ over $A \in \Omega^{+}_{1,b,P}$. Both used, at some points in their proofs, the idea of comparing $cov(A)$ with $cov(A')$, where A' is obtained from A by perturbing just two elements of A. Below, we formalize this idea and show that it can be used as the sole tool to prove both that $cov_{1,b,k}(A)$, $A \in \Omega^{+}_{1,b,P}$, has a global maximum at $A=\left[\frac{p}{b}\right]_{1,b}$, and that $cov_{1,b,k}(A)$, $A \in \Omega^{\mathbf{N}}_{1,b,P}$, is maximized whenever A is the key distribution of an even-sprinkling function.

Intuitively, what we shall call a **simple spreading** is a transformation on an array (in $\Omega_{1,b,P}$) which selects two elements of the array and increases their difference without changing their sum (whence a spreading); and leaves all other elements unchanged (whence simple).

Definition 10: Let $A, A' \in \Omega_{1,b,P}$. Then A can be **simply spread to** A' (denoted $A \twoheadrightarrow A'$) iff for some $\varepsilon > 0$ and some l and m, $0 \le l, m \le b-1$,

$$\begin{aligned} & a_l \le a_m \\ & a'_l = a_l - \varepsilon \\ & a'_m = a_m + \varepsilon \\ \text{and } \forall i \not\in \{l,m\} \quad & a'_i = a_i \end{aligned}$$

Proposition 2: If $A, A' \in \Omega^{+}_{1,b,P}$ and $A \twoheadrightarrow A'$, then $cov_{1,b,k}(A) \ge cov_{1,b,k}(A')$. Furthermore, if $cov_{1,b,k}(A) > 0$, then $cov_{1,b,k}(A) > cov_{1,b,k}(A')$.

Proof: The only terms which differ between $cov_{1,b,k}(A)$ and $cov_{1,b,k}(A')$ are those involving a_l, a'_l, or a_m, a'_m, or both. Hence

$$\begin{aligned} cov_{1,b,k}(A) - cov_{1,b,k}(A') = {} & (a_l - a'_l) \sum_{\substack{I \subseteq b \setminus \{l,m\} \\ |I| = k-1}} \prod_{i \in I} a_i \\ & + (a_m - a'_m) \sum_{\substack{I \subseteq b \setminus \{l,m\} \\ |I| = k-1}} \prod_{i \in I} a_i \\ & + (a_l a_m - a'_l a'_m) \sum_{\substack{I \subseteq b \setminus \{l,m\} \\ |I| = k-2}} \prod_{i \in I} a_i \\ = {} & ((a_l - a'_l) + (a_m - a'_m)) \sum_{\substack{I \subseteq b \setminus \{l,m\} \\ |I| = k-1}} \prod_{i \in I} a_i \\ & + (a_l a_m - a'_l a'_m) \sum_{\substack{I \subseteq b \setminus \{l,m\} \\ |I| = k-2}} \prod_{i \in I} a_i \\ = {} & (\varepsilon + (-\varepsilon)) \sum_{\substack{I \subseteq b \setminus \{l,m\} \\ |I| = k-1}} \prod_{i \in I} a_i \\ & + (a_l a_m - (a_l a_m - \varepsilon a_m + \varepsilon a_l - \varepsilon^2)) \sum_{\substack{I \subseteq b \setminus \{l,m\} \\ |I| = k-2}} \prod_{i \in I} a_i \end{aligned}$$

$$= \varepsilon\,(a_m - a_l + \varepsilon) \sum_{\substack{I \subseteq b\setminus\{l,m\} \\ |I| = k-2}} \prod_{i \in I} a_i$$

Since $\varepsilon > 0$ and $a_m \geq a_l$ and all the a_i are nonnegative, the last expression for the difference is nonnegative. Thus $cov_{1,b,k}(A) \geq cov_{1,b,k}(A')$.

Now suppose that $cov_{1,b,k}(A) > 0$. Then there must exist some subset of k elements of A, all of which are nonzero. Hence, even excluding a_l and a_m, there is still some subset of $k-2$ nonzero elements of A. This means that

$$\sum_{\substack{I \subseteq b\setminus\{l,m\} \\ |I| = k-2}} \prod_{i \in I} a_i > 0.$$

Thus the last expression for the difference is strictly positive and $cov_{1,b,k}(A) > cov_{1,b,k}(A')$. This completes the proof of the proposition.

Corollary: If $A, C \in \Omega^{+}_{1,b,p}$, $A \neq C$, $cov_{1,b,k}(A) > 0$, and there exists a sequence $A = A^0 \twoheadrightarrow A^1 \twoheadrightarrow \ldots \twoheadrightarrow A^n = C$ then $cov_{1,b,k}(A) > cov_{1,b,k}(C)$.

Proposition 3: $cov_{1,b,k}(A)$, $A \in \Omega^{+}_{1,b,p}$ has a global maximum at $A = \left[\frac{p}{b}\right]_{1,b} = [\frac{p}{b}, \ldots, \frac{p}{b}]$.

Proof: We will show that for every $C \in \Omega^{+}_{1,b,p}$, $C \neq \left[\frac{p}{b}\right]_{1,b} \Rightarrow cov_{1,b,k}(\left[\frac{p}{b}\right]_{1,b}) > cov_{1,b,k}(C)$. Since $cov_{1,b,k}(\left[\frac{p}{b}\right]_{1,b}) > 0$, it suffices, by the previous corollary to exhibit a sequence of simple spreadings which transforms $\left[\frac{p}{b}\right]_{1,b}$ into C.

Let $A^0 = \left[\frac{p}{b}\right]_{1,b}$.

In order to define A^{t+1} for $t > 0$, first identify the positions at which A^t is greater than C, and those at which A^t is less than C:

$$\text{let } L_t = \{ i \mid a_i^t > c_i \},$$
$$\text{and } M_t = \{ i \mid a_i^t < c_i \}.$$

Then pick from each set the position at which the difference is the smallest. (And in case of ties, pick the smallest index.) Let

$$l_t = \min\{ j \in L_t \mid \forall i \in L_t,\ a_j^t - c_j \leq a_i^t - c_i \}$$
$$m_t = \min\{ j \in M_t \mid \forall i \in M_t,\ c_j - a_j^t \leq c_i - a_i^t \}$$

Finally, let

$$\varepsilon_t = \min\{ a_{l_t}^t - c_{l_t}, c_{m_t} - a_{m_t}^t \}$$

and define A^{t+1} by

$$\forall i \neq l_t, m_t \quad a_i^{t+1} = a_i^t$$
$$a_{l_t}^{t+1} = a_{l_t}^t - \varepsilon_t$$

$$a_{m_t}^{t+1} = a_{m_t}^{t} + \varepsilon_t$$

By induction on t we can show

$$\begin{gathered}(1)\ \forall i \in L_t,\ \forall j \in M_t\ \ a_i^t \leq a_j^t, \\ \text{and (2) for some } t_0 < b,\ A^{t_0} = C,\end{gathered}$$

the key observation being that adding and subtracting ε_t preserves (1) and forces progress toward (2). From (1) we have in particular that

$$\forall t\ \ a_{l_t}^t \leq \tfrac{P}{b} \leq a_{m_t}^t .$$

So the sequence we have defined is such that

$$\left[\tfrac{P}{b}\right]_{1,b} = A^0 \twoheadrightarrow A^1 \twoheadrightarrow \ldots \twoheadrightarrow A^{t_0} = C.$$

This concludes the proof of the proposition.

<u>Proposition 4:</u> Suppose $H, C \in \Omega_{1,b,P}^{\mathbf{N}}, H \sim_{\mathbf{N}} \left[\tfrac{P}{b}\right]_{1,b}, C \not\sim_{\mathbf{N}} \left[\tfrac{P}{b}\right]_{1,b}$. Then $cov_{1,b,k}(H) > cov_{1,b,k}(C)$.

> <u>Proof:</u> First, if $H \sim_{\mathbf{N}} \left[\tfrac{P}{b}\right]_{1,b}$ and $H' \sim_{\mathbf{N}} \left[\tfrac{P}{b}\right]_{1,b}$, then the elements of H' are just some permutation of the elements of H and so $cov_{1,b,k}(H) = cov_{1,b,k}(H')$. Second, if $C \not\sim_{\mathbf{N}} \left[\tfrac{P}{b}\right]_{1,b}$, then there exists $H' \in \Omega_{1,b,P}^{\mathbf{N}}$ such that (a) $H' \sim_{\mathbf{N}} \left[\tfrac{P}{b}\right]_{1,b}$; and (b) there exists a sequence $H' = A^0 \twoheadrightarrow A^1 \twoheadrightarrow \ldots \twoheadrightarrow A^t = C$. The proof of (b) is very similar to the proof of the analogous claim in Proposition 3. The only substantial difference is in verifying the induction basis for (1), *i.e.*, that for all i and j, $h_i > c_i$ and $h_j < c_j \Rightarrow h_i \leq h_j$.

Although it seems to have little relevance to our original problem, one might want to know whether it is necessary to restrict the domain of $cov_{1,b,k}$ to $\Omega_{1,b,P}^{+}$ in order for $[\frac{P}{b}]_{1,b}$ to be a global maximum. The answer is yes, and there is an easy example. Consider the point $[5, -1, -1]$ in $\Omega_{1,3,3}$.

$$cov_{1,3,3}([5, -1, -1]) = (5)\cdot(-1)\cdot(-1) = 5 > 1 = cov_{1,3,3}([1, 1, 1])$$

In fact $cov_{1,3,3}$ is unbounded if we allow negative entries: for each real number c,

$$\begin{aligned} cov_{1,3,3}([1+2c, 1-c, 1-c]) &= (1+2c)\cdot(1-c)\cdot(1-c) \\ &= 1 - 3c^2 + 2c^3, \end{aligned}$$

which is unbounded in the positive direction as c increases.

The case of two keys per set (k=2)

Berman, Bock, and Plank [Ber 81] have exhibited a construction which produces a set of functions $\{f_1,\ldots,f_n\} \subseteq b^P$ covering all $(2,P)$-sets such that (a) $\{f_1,\ldots,f_n\}$ is of minimal size among the subsets of b^P which cover all $(2,P)$-sets; and (b) every $f \in \{f_1,\ldots,f_n\}$ is an even-sprinkling function. They show further that $\lfloor \log_b P \rfloor \leq n \leq \lceil \log_b P \rceil$.

By showing that the converse of Proposition 1 holds in the case $k=2$, we establish a complete characterization of the subsets of b^P which cover all $(2,P)$-sets, and show that the minimal size for such a subset is exactly $\lceil \log_b P \rceil$.

Proposition 5: $\{f_1,\ldots,f_n\} \subseteq b^P$ covers all $(2,P)$-sets iff $\forall \mu \; a_\mu(\{f_1,\ldots,f_n\}) \in \{0,1\}$.

Proof: ($\Rightarrow$) By Proposition 1.
($\Leftarrow$) Assume that $\forall \mu \; a_\mu(\{f_1,\ldots,f_n\}) \in \{0,1\}$. Pick any $(2,P)$-set, $\{s_1,s_2\}$ and let

$$\mu = (f_1(s_1), f_2(s_1), \ldots, f_n(s_1))$$
$$\text{and } \xi = (f_1(s_2), f_2(s_2), \ldots, f_n(s_2)).$$

Now if $\mu=\xi$, then $a_\mu=a_\xi\geq 2$ since both s_1 and s_2 would be counted there. But this contradicts our hypothesis, so it must be that $\mu\neq\xi$. This implies, of course, that there must be some position, i, at which μ and ξ are not the same ($\mu(i)\neq\xi(i)$). But then

$$f_i(s_1) = \mu(i) \neq \xi(i) = f_i(s_2).$$

Thus f_i covers $\{s_1,s_2\}$. Hence $\{f_1,\ldots,f_n\}$ covers $\{s_1,s_2\}$. Since $\{s_1,s_2\}$ was arbitrary, this completes the proof.

Proposition 6: The smallest n such that there exists $\{f_1,\ldots,f_n\} \subseteq b^P$ which covers all $(2,P)$-sets is $n = \lceil \log_b P \rceil$.

Proof: By Proposition 1, no n smaller than $\lceil \log_b P \rceil$ will do. On the other hand, set $n = \lceil \log_b P \rceil$ and pick any $A \in \Omega^{\mathrm{N}}_{n,b,P}$ such that $\forall \mu \; a_\mu \in \{0,1\}$. Then A is the key distribution of some $F = \{f_1,\ldots,f_n\} \subset b^P$. By Proposition 5, F covers all $(2,P)$-sets.

The preceding results took aim directly at Question 2. The next result answers Question 1 for $k=2$ in the (unlikely) case that $b^n \mid P$. It can be obtained as a corollary to Theorem 1 in the next section, but we insert it here because the proof generates the answer to our question "Where is the maximum?" from scratch, instead of confirming a guess, as the proofs of Proposition 8 and Theorem 1 do.

Proposition 7: For all $n\geq 1, b\geq 2$, and $P>0$, $cov_{n,b,2}(A)$ has a global maximum for $A \in \Omega_{n,b,P}$ at $A=\left[\frac{P}{b^n}\right]_{n,b}$.

Proof: From the definition of $cov_{n,b,k}$, with $k=2$, we have

$$cov_{n,b,2}(A) = \tfrac{1}{2} \sum_{diff_2(\mu,\xi)\geq 1} a_\mu a_\xi .$$

Now note that $diff_2(\mu,\xi)\geq 1$ if and only if $\mu\neq\xi$. Thus

$$cov_{n,b,2}(A) = \tfrac{1}{2}\sum_{\mu} \sum_{\xi\neq\mu} a_\mu a_\xi. \tag{1}$$

First we would like to find all the critical points of this function when its domain is restricted to $\Omega_{n,b,P}$ for arbitrary P. To do this, we employ the method of Lagrange multipliers, viewing the condition $\sum_{\mu} a_\mu = P$ as a side constraint. The augmented function then is

$$G(A) = \tfrac{1}{2}\sum_{\mu} \sum_{\xi\neq\mu} a_\mu a_\xi - \lambda\cdot\left[\sum_{\mu} a_\mu - P\right].$$

The critical points will be all solutions to the system of equations obtained by setting the first partials of G equal to zero. Taking the first partials produces

$$\begin{aligned} &\frac{\partial G}{\partial\lambda} = \sum_{\mu} a_\mu - P \\ \forall\mu\ \ &\frac{\partial G}{\partial a_\mu} = \tfrac{1}{2}\sum_{\xi\neq\mu} a_\xi - \lambda \end{aligned}$$

Setting the first partials equal to zero gives

$$\sum_{\mu} a_\mu - P = 0 \tag{2}$$

$$\text{and}\ \ \forall\mu\ \ \tfrac{1}{2}\sum_{\xi\neq\mu} a_\xi - \lambda = 0 \tag{3}$$

From (2) we get

$$\forall\mu\ \ \sum_{\xi\neq\mu} a_\xi = P - a_\mu.$$

Substituting this in (3) gives

$$\begin{aligned} &\forall\mu \quad && \tfrac{1}{2}(P - a_\mu) - \lambda = 0 \\ \Rightarrow\ &\forall\mu && a_\mu = P - 2\lambda \\ \Rightarrow\ &\forall\mu,\mu' && a_\mu = P - 2\lambda = a_{\mu'} \end{aligned}$$

Since $\sum_{\mu} a_\mu = P$ and there are b^n elements a_μ, it follows that

$$\forall\mu\ \ a_\mu = \tfrac{P}{b^n}$$

Thus the only critical point of $cov_{n,b,2}(A)$ for $A \in \Omega_{n,b,P}$ occurs at $A = \left[\frac{P}{b^n}\right]$.

A review of the nearest text on multivariate calculus reveals that in order to determine the nature of a critical point, $\vec{z} = \langle z_1,\dots,z_m\rangle$, of a function of several variables, $f(x_1,\dots,x_m)$, we must know all the second partial derivatives at the point and determine therefrom how the first derivative changes as we move away from the critical point in all the many possible directions. If it decreases in all directions, then we have a strict local maximum; if it increases, a strict local minimum. If it decreases in some directions and increases in others, the situation must be investigated further. These conditions can be conveniently phrased if we arrange the second partials of the function in an $m\times m$ matrix, Q, such that

$$q_{ij} = \frac{\partial^2 f}{\partial z_i \partial z_j}\Big|_{\vec{x}=\vec{z}}$$

Then $\vec{z}$ is a strict local maximum if and only if Q is negative definite, and $\vec{z}$ is a strict local minimum if and only if Q is positive definite. Thus it will suffice to show that the matrix of second partials of $cov_{n,b,2}(A)$ for $A \in \Omega_{n,b,P}$ is negative definite at $A = \left[\frac{P}{b^n}\right]$.

In order to investigate the second partials of $cov_{n,b,2}(A), A \in \Omega_{n,b,P}$ we introduce the constraint on the sum of the elements of A explicitly by picking some μ_0 and substituting $a_{\mu_0} = P - \sum\limits_{\mu \neq \mu_0} a_\mu$ in (1). First from (1) we separate out the terms containing a_{μ_0}:

$$\begin{aligned}
cov_{n,b,2}(A) &= \tfrac{1}{2}\sum_{\xi\neq\mu_0} a_{\mu_0}a_\xi \;+\; \tfrac{1}{2}\sum_{\mu\neq\mu_0}\sum_{\xi\neq\mu} a_\mu a_\xi \\
&= \tfrac{1}{2}\sum_{\xi\neq\mu_0} a_{\mu_0}a_\xi \;+\; \tfrac{1}{2}\sum_{\mu\neq\mu_0}\left[a_\mu a_{\mu_0} + \sum_{\xi\neq\mu,\mu_0} a_\mu a_\xi\right] \\
&= \tfrac{1}{2}\sum_{\xi\neq\mu_0} a_{\mu_0}a_\xi \;+\; \tfrac{1}{2}\sum_{\mu\neq\mu_0} a_\mu a_{\mu_0} \;+\; \tfrac{1}{2}\sum_{\mu\neq\mu_0}\sum_{\xi\neq\mu,\mu_0} a_\mu a_\xi \\
&= \sum_{\xi\neq\mu_0} a_{\mu_0}a_\xi \;+\; \tfrac{1}{2}\sum_{\mu\neq\mu_0}\sum_{\xi\neq\mu,\mu_0} a_\mu a_\xi
\end{aligned}$$

Then we substitute for a_{μ_0}

$$\begin{aligned}
cov_{n,b,2}(A) &= \sum_{\xi\neq\mu_0} a_\xi\,(P - \sum_{\mu\neq\mu_0} a_\mu) \;+\; \tfrac{1}{2}\sum_{\mu\neq\mu_0}\sum_{\xi\neq\mu,\mu_0} a_\mu a_\xi \\
&= P\sum_{\xi\neq\mu_0} a_\xi \;-\; \sum_{\xi\neq\mu_0}\sum_{\mu\neq\mu_0} a_\xi a_\mu \;+\; \tfrac{1}{2}\sum_{\mu\neq\mu_0}\sum_{\xi\neq\mu,\mu_0} a_\mu a_\xi
\end{aligned}$$

Thus

$$\begin{aligned}
&\forall\mu\neq\mu_0,\ \xi\neq\mu,\mu_0, && \frac{\partial^2 cov_{n,b,2}}{\partial a_\mu \partial a_\xi} = -1 + \tfrac{1}{2} = -\tfrac{1}{2} \\
&\text{and } \forall\mu\neq\mu_0 && \frac{\partial^2 cov_{n,b,2}}{\partial a_\mu^2} = -1
\end{aligned}$$

So the matrix of second partials, Q, is a $(b^n-1 \times b^n-1)$ matrix with -1 on the diagonal, and $-\frac{1}{2}$ everywhere else.

To see that this matrix is negative definite recall that a matrix is negative definite if all its eigenvalues are negative. Furthermore, the eigenvalues of a matrix, M, are just those values λ which make $M-\lambda I$ singular, and the multiplicity of an eigenvalue, λ, is the dimension of the kernel of $M-\lambda I$. Clearly, $-\frac{1}{2}$ is an eigenvalue of Q of multiplicity b^n-2. It is only a little less obvious that $-\frac{b^n}{2}$ is an eigenvalue of Q of multiplicity 1. The sum of their multiplicities being b^n-1, which is the order of Q, these are all the eigenvalues of Q, and clearly they are both negative. Therefore Q is negative definite, and hence $A = \left[\frac{P}{b^n}\right]$ is a strict local maximum. Being the only critical point, it is also a global maximum. This completes the proof of the proposition.

This suggests that when looking for maxima of $cov_{n,b,k}(A)$ (a) for $A \in \Omega^+_{n,b,P}$, $A = \left[\frac{P}{b^n}\right]_{n,b}$ may be a good place to look; (b) for $A \in \Omega^{\mathbf{N}}_{n,b,P}$, points close to $\left[\frac{P}{b^n}\right]_{n,b}$ may be good candidates.

For $k=2$, at least, (b) bears out well.

Proposition 8: For $A \in \Omega^{\mathbf{N}}_{n,b,P}$, $cov_{n,b,2}(A)$ is maximized whenever $A \sim_{\mathbf{N}} \left[\frac{P}{b^n}\right]_{n,b}$.

> Proof: By a "finite-series-of-perturbations" argument like that used to prove the results for $cov_{1,b,k}$ in the previous section.

The general case

If we have more than one function and more than two keys per set, the only result so far which applies is Proposition 1 (and it addresses only Question 2). The theorem in this section represents the extent of our other progress in the general case.

We are looking for maxima of $cov_{n,b,k}(A), A \in \Omega_{n,b,P}$. A preliminary observation will simplify the notation which follows. Since $cov_{n,b,k}(A)$ is the sum of products of length k, it follows that for any positive constant c

$$cov_{n,b,k}(c \cdot A) = c^k \cdot cov_{n,b,k}(A).$$

This implies that so as long as we are operating in the realm of positive real numbers, if we find a maximum for any particular value of P, then we can easily find corresponding maxima for all other values of P. Being free to pick a convenient value of P to work with, we will choose $P=b^n$, so that $\left[\frac{P}{b^n}\right]_{n,b} = [1]_{n,b}$.

Theorem 1: $\forall n \geq 1$, $\forall k \geq 2$, $\forall b \geq k$, $cov_{n,b,k}(A)$ has a strict local maximum for $A \in \Omega_{n,b,b^n}$ at $A=[1]_{n,b}$.

Proof: First we will give a short sketch of the proof. We will try to learn something of the behavior of $cov_{n,b,k}(A)$ near the point $A = [1]_{n,b}$ by examining the behavior of $cov_{n,b,k}$ along lines in Ω_{n,b,b^n} which pass through $[1]_{n,b}$. If D is an n-dimensional $b \times, \ldots, \times b$ array, the elements of which sum to 0 ($D \in \Omega_{n,b,0}$), but which has nonzero elements ($D \neq [0]_{n,b}$), then $\{(\rho \cdot D + [1]_{n,b}) \mid \rho \in \mathbf{R}\}$ is such a line. Furthermore every line in Ω_{n,b,b^n} which passes through $[1]_{n,b}$ can be characterized in this way for some $D \in \Omega_{n,b,0} \setminus \{[0]_{n,b}\}$.

Thinking of D as fixed, we see that $cov_{n,b,k}(\rho D + [1]_{n,b})$ is a function of the single variable ρ. Our first step will be to show that in fact $cov_{n,b,k}(\rho D + [1]_{n,b})$ is a polynomial in ρ of degree k, *i.e.*, there are coefficients $c_0, \ldots, c_k$ such that

$$cov_{n,b,k}(\rho D + [1]_{n,b}) = \sum_{i=0}^{k} c_k \cdot \rho^k.$$

Of course, the coefficients depend on the choice of D, so we write

$$cov_{n,b,k}(\rho D + [1]_{n,b}) = \sum_{i=0}^{k} c_k(D) \cdot \rho^k.$$

The next portion of the proof will demonstrate that $\forall D \in \Omega_{n,b,0} \backslash \{[0]_{n,b}\}$

$$\begin{aligned} c_0(D) &= cov_{n,b,k}([1]_{n,b}), \\ c_1(D) &= 0, \\ \text{and } c_2(D) &< 0. \end{aligned}$$

These facts will then be used to show that $A = [1]_{n,b}$ is a strict local maximum.

And now we present the proof in detail.
From the definition of $cov_{n,b,k}$ we have that for any $A, B \in \bigcup_{P \geq 0} \Omega_{n,b,P}$,

$$cov_{n,b,k}(B+A) = \tfrac{1}{k!} \sum_{diff_k(\xi_1,\ldots,\xi_k) \geq 1} \prod_{j=1}^{k} (b_{\xi_j} + a_{\xi_j}).$$

The idea for the following expansion comes from a similar expansion in [Sas 69]. Each product $\prod_{j=1}^{k} (b_{\xi_j} + a_{\xi_j})$ can be multiplied out into 2^k terms, each the product of some elements from B and/or some elements from A. By expanding all the products in this way and then rearranging the factors within each of the resulting terms we obtain an expression for $cov_{n,b,k}(B+A)$ consisting of terms of the form

$$\prod_{j=1}^{r} b_{\mu_j} \cdot \prod_{j=r+1}^{k} a_{\mu_j}$$

for various values of r and various sequences $\mu_1,\ldots,\mu_k$. Now it is clear that the term $\prod_{j=1}^{r} b_{\mu_j} \cdot \prod_{j=r+1}^{k} a_{\mu_j}$ appears if and only if $\mu_1,\ldots,\mu_k$ is some permutation of some $\xi_1,\ldots,\xi_k$ such that $diff_k(\xi_1,\ldots,\xi_k) \geq 1$; hence, if and only if $diff_k(\mu_1,\ldots,\mu_k) \geq 1$. So all we have left to do is a little counting.

For a particular value of r and sequence $\mu_1,\ldots,\mu_k$, a product $\prod_{j=1}^{k} (b_{\xi_j} + a_{\xi_j})$ in the original expression will produce one term $\prod_{j=1}^{r} b_{\mu_j} \cdot \prod_{j=r+1}^{k} a_{\mu_j}$ if and only if $(\mu_1,\ldots,\mu_r)$ and $(\mu_{r+1},\ldots,\mu_k)$ are both subsequences of $\xi_1,\ldots,\xi_k$. Since there are $\binom{k}{r} = \frac{k!}{r!(k-r)!}$ sequences $\xi_1,\ldots,\xi_k$ which meet this condition we have

$$cov_{n,b,k}(B+A) = \tfrac{1}{k!} \sum_{diff_k(\mu_1,\ldots,\mu_k) \geq 1} \sum_{r=0}^{k} \tfrac{k!}{r!(k-r)!} \left[\prod_{j=1}^{r} b_{\mu_j} \right] \left[\prod_{j=r+1}^{k} a_{\mu_j} \right].$$

The factors $\frac{1}{k!}$ and $k!$ cancel, and since the summations are independent we can reverse their order, yielding

$$cov_{n,b,k}(B+A) = \sum_{r=0}^{k} \sum_{diff_k(\mu_1,\ldots,\mu_k) \geq 1} \tfrac{1}{r!(k-r)!} \left[\prod_{j=1}^{r} b_{\mu_j} \right] \left[\prod_{j=r+1}^{k} a_{\mu_j} \right]$$

We pause here to make an observation which will be useful later. The term of the preceding expression obtained when $r=0$ is

$$\sum_{\mathit{diff}_k(\mu_1,\ldots,\mu_k)\geq 1} \tfrac{1}{k!}\prod_{j=1}^{k} a_{\mu_j} = cov_{n,b,k}(A)$$

no matter what B is. Leaving this aside for the moment and substituting ρD for B, and $[1]_{n,b}$ for A we have

$$\begin{aligned} cov_{n,b,k}(\rho D+[1]_{n,b}) &= \sum_{r=0}^{k} \sum_{\mathit{diff}_k(\mu_1,\ldots,\mu_k)\geq 1} \tfrac{1}{r!(k-r)!}\left[\prod_{j=1}^{r}\rho d_{\mu_j}\right]\left[\prod_{j=r+1}^{k} 1\right] \\ &= \sum_{r=0}^{k} \sum_{\mathit{diff}_k(\mu_1,\ldots,\mu_k)\geq 1} \tfrac{1}{r!(k-r)!}\rho^r \prod_{j=1}^{r} d_{\mu_j} \\ &= \sum_{r=0}^{k} \rho^r\cdot\tfrac{1}{r!(k-r)!} \sum_{\mathit{diff}_k(\mu_1,\ldots,\mu_k)\geq 1}\prod_{j=1}^{r} d_{\mu_j} \end{aligned}$$

Note that although the second sum is over sequences $\mu_1,\ldots,\mu_k$, only $\mu_1,\ldots,\mu_r$ actually appear in the subsequent expression. In order to change the summation to be over just $\mu_1,\ldots,\mu_r$, we must determine, given $\mu_1,\ldots,\mu_r$, how many sequences $\mu_{r+1},\ldots,\mu_k$ there are such that $\mathit{diff}_k(\mu_1,\ldots,\mu_k)\geq 1$. We can visualize the problem as follows:

$$\begin{array}{ccc} \mu_1(1) & \cdots & \mu_1(n) \\ \mu_2(1) & \cdots & \mu_2(n) \\ \vdots & & \vdots \\ \mu_r(1) & \cdots & \mu_r(n) \\ \hline \mu_{r+1}(1) & \cdots & \mu_{r+1}(n) \\ \vdots & & \vdots \\ \mu_k(1) & \cdots & \mu_k(n) \end{array}$$

Given values above the line we must count the number of ways of filling in values (from $\{0,\ldots,b-1\}$) below the line so that at least one entire column is "good" (comprises k distinct integers).

Suppose that exactly e columns above the line are good (*i.e.*, $\mathit{diff}_r(\mu_1,\ldots,\mu_r)=e$). W.l.o.g. we may assume that these are columns $1,\ldots,e$. Clearly, columns which are "bad" (not good) above the line will be bad as entire columns no matter how we extend them by filling in below the line. Thus, whether a particular choice of values for entries below the line satisfies the condition of producing at least one good entire column depends only on how it extends columns $1,\ldots,e$ below the line. To extend a column good above the line to a good entire column we must choose $k-r$ entries from the $b-r$ integers not used above the line. Since each such choice may be ordered in $(k-r)!$ ways, there are

$$\tbinom{b-r}{k-r}\cdot(k-r)! = \tfrac{(b-r)!}{(b-k)!}$$

good extensions of a column good above the line. There being b^{k-r} ways to fill in a column below the line we conclude that there are

$$b^{k-r} - \tfrac{(b-r)!}{(b-k)!}$$

bad extensions of a column good above the line.

Now, in order ***not*** to satisfy the condition requiring one good column, we must choose a bad extension for each of columns $1,\dots,e$. Since the choices are independent, we see that there are

$$(b^{k-r} - \tfrac{(b-r)!}{(b-k)!})^e$$

ways to do this. Since there are $b^{(k-r)e}$ ways in all to extend columns $1,\dots,e$, the number of ways to produce at least one good column must be

$$b^{(k-r)e} - (b^{k-r} - \tfrac{(b-r)!}{(b-k)!})^e .$$

Recalling that we can extend columns $e+1,\dots,n$ in any of the $b^{(n-e)(k-r)}$ possible ways, the final count is

$$\begin{aligned} & b^{(n-e)(k-r)}\cdot\left[b^{(k-r)e} - (b^{k-r} - \tfrac{(b-r)!}{(b-k)!})^e\right] \\ = {} & b^{n(k-r)}\cdot\left[1 - b^{-(k-r)e}\cdot(b^{k-r} - \tfrac{(b-r)!}{(b-k)!})^e\right] \\ = {} & b^{n(k-r)}\cdot\left[1 - (b^{-(k-r)}\cdot(b^{k-r} - \tfrac{(b-r)!}{(b-k)!}))^e\right] \\ = {} & b^{n(k-r)}\cdot\left[1 - (1 - \tfrac{(b-r)!}{(b-k)!\,b^{k-r}})^e\right]. \end{aligned}$$

Continuing our massaging of $cov_{n,b,k}$ we can now transform the second summation as we wanted to, yielding

$$\begin{aligned} cov_{n,b,k}(\rho D+[1]_{n,b}) &= \sum_{r=0}^{k} \rho^r \cdot \tfrac{1}{r!(k-r)!} \sum_{\mu_1,\dots,\mu_r} b^{n(k-r)}\cdot\left[1 - (1 - \tfrac{(b-r)!}{(b-k)!\,b^{k-r}})^{diff_r(\mu_1,\dots,\mu_r)}\right]\cdot\prod_{j=1}^{r} d_{\mu_j} \\ &= \sum_{r=0}^{k} \rho^r \cdot \tfrac{b^{n(k-r)}}{r!(k-r)!} \sum_{\mu_1,\dots,\mu_r} (1 - w(b,k,r)^{diff_r(\mu_1,\dots,\mu_r)}) \prod_{j=1}^{r} d_{\mu_j} \end{aligned}$$

where

$$w(b,k,r) = (1 - \tfrac{(b-r)!}{(b-k)!\,b^{k-r}})$$

(One might expect that the second summation would be over "$diff_r(\mu_1,\dots,\mu_r) \geq 1$". In fact, that would be an equivalent expression to the one above, because the result of our counting exercise has taken care of this: $(1-w(b,k,r)^{diff_r(\mu_1,\dots,\mu_r)})$ evaluates to 0 if $diff_r(\mu_1,\dots,\mu_r)=0$.) For a fixed D the expression above is clearly a polynomial in ρ, with coefficients

$$c_r(D) = \tfrac{b^{n(k-r)}}{r!(k-r)!} \sum_{\mu_1,\dots,\mu_r} (1 - w(b,k,r)^{diff_r(\mu_1,\dots,\mu_r)}) \prod_{j=1}^{r} d_{\mu_j}.$$

Next we investigate the behavior of these coefficients.

Recall the observation made about the expanded expression for $cov_{n,b,k}(B+A)$, that the term obtained when $r=0$ is just $cov_{n,b,k}(A)$. In light of the fact that we have substituted $[1]_{n,b}$ for A, this implies that for all $D \in \Omega_{n,b,0}$

$$c_0(D) = cov_{n,b,k}([1]_{n,b}).$$

For the other coefficients, note that if the factor $(1-w(b,k,r)^{diff_r(\mu_1,\dots,\mu_r)})$ could be moved outside the summation without muddying the waters too much, then the expression would become rather simple. For most values of r this is not possible, because $\mathit{diff}_r(\mu_1,\ldots,\mu_r)$ depends in a complicated manner on $\mu_1,\ldots,\mu_r$. However, for $r=1$ and $r=2$, $\mathit{diff}_r(\mu_1,\ldots,\mu_r)$ behaves reasonably well. In particular, for $r=1$, $\forall\mu$, $\mathit{diff}_1(\mu)=n$. So $\forall D\in\Omega_{n,b,0}$

$$\begin{aligned}c_1(D) &= \frac{b^{n(k-1)}}{(k-1)!}\sum_{\mu}(1-w(b,k,1)^n)\cdot d_\mu\\ &= \frac{b^{n(k-1)}\cdot(1-w(b,k,1)^n)}{(k-1)!}\cdot\sum_{\mu}d_\mu\\ &= 0\end{aligned}$$

since the sum of the elements of D is 0 by hypothesis.

In the case of the coefficient of ρ^2 we have

$$\begin{aligned}c_2(D) &= \frac{b^{n(k-2)}}{2\cdot(k-2)!}\sum_{\mu}\sum_{\xi}d_\mu\cdot d_\xi\cdot(1-w(b,k,2)^{\mathit{diff}_2(\mu,\xi)})\\ &= \frac{b^{n(k-2)}}{2\cdot(k-2)!}\left[\sum_{\mu}\sum_{\xi}d_\mu\cdot d_\xi-\sum_{\mu}\sum_{\xi}d_\mu\cdot d_\xi\cdot w(b,k,2)^{\mathit{diff}_2(\mu,\xi)}\right]\\ &= \frac{b^{n(k-2)}}{2\cdot(k-2)!}\left[\left(\sum_{\mu}d_\mu\right)^2-\sum_{\mu}\sum_{\xi}d_\mu\cdot d_\xi\cdot w(b,k,2)^{\mathit{diff}_2(\mu,\xi)}\right]\end{aligned}$$

Since $\sum_{\mu}d_\mu=0$ this simplifies to

$$c_2(D) = \frac{b^{n(k-2)}}{2\cdot(k-2)!}\left[-\sum_{\mu}\sum_{\xi}d_\mu\cdot d_\xi\cdot w(b,k,2)^{\mathit{diff}_2(\mu,\xi)}\right]$$

Now we need to know a little more about $w(b,k,2)$. For all b,k,r, $b\geq2$, $1\leq k\leq b$, $0\leq r\leq k$,

$$\begin{aligned}w(b,k,r) &= (1-\frac{(b-r)!}{(b-k)!b^{k-r}})\\ &= (1-\frac{(b-r)\cdot(b-r-1)\cdot\ldots\cdot(b-k+1)}{b^{k-r}})\\ &= (1-(\frac{b-r}{b}\cdots\frac{b-k+1}{b}))\end{aligned}$$

Since each of the fractions in the second term is in (0,1], their product is also in (0,1]. Hence

$$0\leq w(b,k,r)<1.$$

Given this, and $D\neq[0]_{n,b}$, the lemma which follows this proof shows that

$$\sum_{\mu}\sum_{\xi}d_\mu\cdot d_\xi\cdot w(b,k,2)^{\mathit{diff}_2(\mu,\xi)}>0.$$

Therefore $\forall D\in\Omega_{n,b,0}\setminus\{[0]_{n,b}\}$

$$c_2(D) = \frac{b^{n(k-2)}}{2\cdot(k-2)!}\left[- \sum_{\mu}\sum_{\xi} d_{\mu}\cdot d_{\xi}\cdot w(b,k,2)^{diff_2(\mu,\xi)}\right] < 0.$$

Thus we have

$$cov_{n,b,k}(\rho D+[1]_{n,b}) = cov_{n,b,k}([1]_{n,b}) + c_2(D)\rho^2$$
$$+ \text{ terms in higher powers of } \rho.$$

For sufficiently small values of ρ, the ρ^2 term will dominate. Since $c_2(D)$ is negative, we have that for sufficiently small ρ

$$cov_{n,b,k}(\rho D+[1]_{n,b}) < cov_{n,b,k}([1]_{n,b}).$$

Hence, $cov_{n,b,k}([1]_{n,b})$ is a strict local maximum. This completes the proof of the theorem.

Lemma: If $b \geq 1$, z is a constant, $0 \leq z < 1$, and

$$H(A) = \sum_{\mu}\sum_{\xi} a_{\mu}\cdot a_{\xi}\cdot z^{diff_2(\mu,\xi)},$$

then $\forall n \geq 0$, $\forall A \in \bigcup_{P} \Omega_{n,b,P}$, $H(A) \geq 0$, with equality iff $A=[0]_{n,b}$.

Proof: By induction on n.

$n=0$

In this case A has but one element, a, and

$$H(A) = \sum_{\mu}\sum_{\xi} a_{\mu}\cdot a_{\xi}\cdot z^{diff_2(\mu,\xi)}$$
$$= a^2\cdot z^0$$
$$= a^2$$

So $H(A) \geq 0$ in all cases, and $H(A) = 0$ if and only if $A = [0]$.

$n>0$

Assume that $H(A) \geq 0$ for $A \in \bigcup_{P} \Omega_{n-1,b,P}$, with equality if and only if $A = [0]_{n-1,b}$, and now consider $A \in \Omega_{n,b,P}$.

First we need a notation which will allow us to pass easily from indices for elements of $n-1$-dimensional arrays to indices for elements of A (an n-dimensional array). To this end, for $\mu : n-1 \to b$ and $i \in b$, let

$$a_{(\mu,i)} = a_{\mu(0),\mu(1),\ldots,\mu(n-2),i}.$$

Now define $S \in \Omega_{n-1,b,P}$ by

$$s_\mu = \sum_{i=0}^{b-1} a_{(\mu,i)}.$$

Also, for $0 \le i \le b-1$, define A^i by

$$a^i{}_\mu = a_{(\mu,i)}.$$

S, then, is a condensed version of A, obtained by summing elements having the same first $n-1$ coordinates, and A^i is a slice of A. The idea of the main part of the proof is that $H(S)$ contains exactly the products $a_{(\mu,i)}a_{(\xi,j)}$ that $H(A)$ contains. The difference is that in $H(A)$ the product $a_{(\mu,i)}a_{(\xi,j)}$ is multiplied by $z^{diff_2((\mu,i),(\xi,j))}$, whereas in $H(S)$ it is multiplied by $z^{diff_2(\mu,\xi)}$. These are the same if $i=j$, but differ by a factor of z if $i \ne j$. Fortunately, this difference can be easily expressed in terms of $H(A^i)$ and $H(A^j)$.

Formally, we can write $H(A)$ as

$$H(A) = \sum_\mu \sum_\xi \sum_i \sum_j a_{(\mu,i)}a_{(\xi,j)}z^{diff_2((\mu,i),(\xi,j))}.$$

Now we separate the cases in which $i=j$ from those in which $i \ne j$ and use our earlier observation that $z^{diff_2((\mu,i),(\xi,j))} = \begin{cases} z^{diff_2(\mu,\xi)}, & \text{if } i=j \\ z \cdot z^{diff_2(\mu,\xi)}, & \text{if } i \ne j \end{cases}$

$$\begin{aligned} H(A) &= \sum_i \sum_{j \ne i} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2((\mu,i),(\xi,j))} + \sum_{i=j} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2((\mu,i),(\xi,j))} \\ &= \sum_i \sum_{j \ne i} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)} \cdot z \cdot z^{diff_2(\mu,\xi)} + \sum_{i=j} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2(\mu,\xi)} \\ &= z \cdot \sum_i \sum_{j \ne i} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2(\mu,\xi)} + \sum_{i=j} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2(\mu,\xi)} \end{aligned}$$

The next step is to add and subtract just what we need to make the first term into $H(S)$.

$$\begin{aligned} H(A) = z \cdot \sum_i \sum_{j \ne i} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2(\mu,\xi)} &+ z \cdot \sum_{i=j} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2(\mu,\xi)} \\ - z \cdot \sum_{i=j} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2(\mu,\xi)} &+ \sum_{i=j} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2(\mu,\xi)} \\ = z \cdot \sum_i \sum_j \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2(\mu,\xi)} &+ (1-z) \cdot \sum_{i=j} \sum_\mu \sum_\xi a_{(\mu,i)}a_{(\xi,j)}z^{diff_2(\mu,\xi)} \end{aligned}$$

$$= z \cdot \sum_{\mu} \sum_{\xi} \sum_{i} \sum_{j} a_{(\mu,i)} a_{(\xi,j)} z^{\mathit{diff}_2(\mu,\xi)} + (1-z) \cdot \sum_{i} \sum_{\mu} \sum_{\xi} a_{(\mu,i)} a_{(\xi,i)} z^{\mathit{diff}_2(\mu,\xi)}$$

$$= z \cdot \sum_{\mu} \sum_{\xi} \left[\sum_{i} a_{(\mu,i)}\right] \left[\sum_{j} a_{(\xi,j)}\right] z^{\mathit{diff}_2(\mu,\xi)} + (1-z) \cdot \sum_{i} \sum_{\mu} \sum_{\xi} a^{i}{}_{\mu} a^{i}{}_{\xi} z^{\mathit{diff}_2(\mu,\xi)}$$

$$= z \cdot \sum_{\mu} \sum_{\xi} s_{\mu} s_{\xi} z^{\mathit{diff}_2(\mu,\xi)} + (1-z) \cdot \sum_{i} H(A^{i})$$

$$= z\, H(S) + (1-z) \sum_{i} H(A^{i})$$

By the induction hypothesis $H(S) > 0$ and for all i, $H(A^{i}) > 0$. Furthermore, z is nonnegative and, hence, so is $(1-z)$. Thus the expression above is clearly nonnegative (*i.e.*, $H(A) \geq 0$).

Now if $A = [0]_{n,b}$, obviously $H(A) = 0$. On the other hand, suppose $A \neq [0]_{n,b}$, then for some i_0, $A^{i_0} \neq [0]_{n-1,b}$. By the induction hypothesis $H(A^{i_0}) > 0$. Also, $z < 1$, so $(1-z) > 0$. Thus the last expression for $H(A)$ contains at least one strictly positive term. Since it contains no negative terms, this shows that $H(A) > 0$ whenever $A \neq [0]_{n,b}$.

This completes the proof of the lemma.

Corollary (to Theorem 1): $\forall n \geq 1$, $\forall k \geq 2$, $\forall b \geq k$, $\forall P > 0$, $cov_{n,b,k}(A)$, $A \in \Omega_{n,b,P}$ has a strict local maximum at $A = \left[\frac{P}{b^n}\right]_{n,b}$.

For purposes of speculation, we now calculate $cov_{n,b,k}(\left[\frac{P}{b^n}\right]_{n,b})$.

$$cov_{n,b,k}(\left[\frac{P}{b^n}\right]_{n,b}) = \frac{1}{k!} \sum_{\mathit{diff}_k(\mu_1,\ldots,\mu_k) \geq 1} \prod_{j=1}^{k} \frac{P}{b^n}$$

$$= \frac{1}{k!} \left[\frac{P}{b^n}\right]^k \sum_{\mathit{diff}_k(\mu_1,\ldots,\mu_k) \geq 1} 1.$$

A special case ($r = 0$) of the counting exercise which we went through in the proof of Theorem 1 reveals that

$$\sum_{\mathit{diff}_k(\mu_1,\ldots,\mu_k) \geq 1} 1 = b^{nk} \cdot \left[1 - \left(1 - \frac{b!}{(b-k)!\, b^k}\right)^n\right]$$

Hence

$$cov_{n,b,k}(\left[\frac{P}{b^n}\right]_{n,b}) = \frac{1}{k!} \left[\frac{P}{b^n}\right]^k b^{nk} \cdot \left[1 - \left(1 - \frac{b!}{(b-k)!\, b^k}\right)^n\right]$$

$$= \frac{P^k}{k!}\left[1 - (1 - \frac{b!}{(b-k)!b^k})^n\right] \qquad (4)$$

If $A = \left[\frac{P}{b^n}\right]_{n,b}$ is indeed the global maximum for $cov_{n,b,k}(A)$, $A \in \Omega^{+}_{n,b,P}$, then (4) is an upper bound on the number of (k,P)-sets that can be covered by n functions from b^P. Solving

$$cov_{n,b,k}(\left[\frac{P}{b^n}\right]_{n,b}) = \text{total \# of } (k,P)\text{-sets} = \binom{P}{k} = \frac{P!}{(P-k)!k!}$$

for n would then produce a lower bound on the number of functions needed to cover all (k,P)-sets. The outcome of this calculation is

$$n = \frac{\log\left[1 - \frac{P!}{(P-k)!P^k}\right]}{\log\left[1 - \frac{b!}{(b-k)!b^k}\right]}.$$

On the other hand, we would really not expect this to be a tight bound since the key distribution of a set of functions which covered all (k,P)-sets would contain (many) zeroes, and hence would be on the boundary of $\Omega^{+}_{n,b,P}$, whereas $\left[\frac{P}{b^n}\right]_{n,b}$ is right in the middle of $\Omega^{+}_{n,b,P}$.

Summary

In two special cases there are complete answers to Question 1. If we have just one function ($n=1$), or we have just two keys per set ($k=2$), then a set of functions maximizes the number of (k,P)-sets covered if and only if its key distribution is as even as possible, *i.e.*, the elements of its key distribution are equal to $\left\lfloor\frac{P}{b^n}\right\rfloor$ or $\left\lceil\frac{P}{b^n}\right\rceil$ in the correct proportion so as to sum to P. In the case of one function, the functions satisfying this condition are exactly the even-sprinkling functions. In the case of two keys per set (and more than one function) there is always a set of even-sprinkling functions satisfying the condition, though other sets not comprising solely even-sprinkling functions may also satisfy the condition.

As regards Question 2, for the case $k=2$ we know that any set of n functions, the key distribution of which comprises zeroes and ones covers all the $(2,P)$-sets. Since $\lceil\log_b P\rceil$ functions are necessary and sufficient to produce such a key distribution, $\lceil\log_b P\rceil$ is the minimum number of functions which can cover all the $(2,P)$-sets. Furthermore we know that $\lceil\log_b P\rceil$ is a lower bound on the number of functions needed to cover all the (k,P)-sets for k greater than 2.

In the hope of obtaining some approximate answers we have extended the domain of $cov_{n,b,k}(A)$ by allowing any nonnegative real numbers as entries in A, reasoning that a global maximum in this domain would be an upper bound on maxima over the domain of actual key distributions. We have not been able to pin down a global maximum for $cov_{n,b,k}(A)$, $A \in \Omega^{+}_{n,b,P}$ in general, but we have shown that $A = \left[\frac{P}{b^n}\right]_{n,b}$ is a global maximum in case $n=1$ or $k=2$, and is a strict local maximum in all cases. Furthermore, so far we have not found a counterexample to $A = \left[\frac{P}{b^n}\right]_{n,b}$ being a global maximum in general. Based on the conjecture that $A = \left[\frac{P}{b^n}\right]_{n,b}$ is a global maximum in general we have calculated a conjectured upper bound on the number of (k,P)-sets that n functions from b^P can cover, and the corresponding lower bound on the number of functions from b^P needed to cover all (k,P)-sets.

References

[And 79] Anderson, M. and Sprugnoli, R. Unpublished notes.

[Ber 81] Berman, F., Bock, M., and Plank, D. Unpublished notes.

[Car 77] Carter, J.L. and Wegman, M.N. Universal Classes of Hash Functions. *Proceedings of the Ninth Annual Symposium on the Theory of Computing*, May 1977 p. 106-112

[Cic 80] Cichelli, R. Minimal Perfect Hash Functions Made Simple. *CACM 23*, 1 (January 1980) p. 17-19

[Sas 69] Sasser, D.W. and Slater, M.L. On a Generalization of the Van der Waerden Conjecture. *Portugaliae Mathematica 28* (1969) p. 91-95

[Spr 77] Sprugnoli, R. Perfect Hashing Functions: A Single Probe Retrieving Method for Static Sets. *CACM 20*, 11 (November 1977) p. 841-850

[Tar 79] Tarjan, R.E. and Yao, A.C. Storing a Sparse Table. *CACM 22*, 11 (November 1979) p. 606-611