Pairwise-Independent Hashing – Extra Notes

The following notes complement the class slides. In particular, here, we are interested in the following problem: We have a set $S = \{s_1, \ldots, s_n\}$ of n elements, and $S \subseteq [K]$ – i.e., S is a subset of the integers from 1 to K. Now, we pick a function $\mathbf{h} : [K] \to [M]$ from the set of all such functions, uniformly at random (i.e., all of them are equally likely to be picked), and are interested in the event that there exists a *collision*, i.e., two distinct elements of S – call them s and s' – which are distinct *and* such that $\mathbf{h}(s) = \mathbf{h}(s')$. Below, we will explain how to achieve the same result by sampling $\mathbf{h}$ from a small set of functions.

Setting up the problem. We can define this as a probability space as follows. We let $\Omega = \mathcal{H}_{K,M}$, where $\mathcal{H}_{K,M}$ is the set of *all* functions from $[K]$ to $[M]$ – in particular, $|\mathcal{H}_{K,M}| = M^K$. We also have that all functions $\mathbf{h}$ are equally likely to be picked, i.e., $\mathbb{P}(\mathbf{h}) = 1/M^K$ for all $\mathbf{h} \in \mathcal{H}_{K,M}$.

Then, the following event is the event that a collision occurs, i.e.,

$$\mathcal{C} = \left\{\mathbf{h} : \exists s, s' \in S : s \neq s' \wedge \mathbf{h}(s) = \mathbf{h}(s')\right\} .$$

We now prove the following.

Theorem 1. $\mathbb{P}(\mathcal{C}) \leqslant \frac{n(n-1)}{2M}$.

Proof. Assume for simplicity $S = \{1, \ldots, n\}$. The argument will not depend on this, but this makes the notation simpler. The first thing we want to do is to rewrite the event $\mathcal{C}$ as the union of smaller events. In particular, we let $\mathcal{C}_{i,j}$ be the event (for $1 \leqslant i < j \leqslant n$) that $\mathbf{h}(i) = \mathbf{h}(j)$, i.e.,

$$\mathcal{C}_{i,j} = \{\mathbf{h} \in \mathcal{H}_{K,M} : \mathbf{h}(i) = \mathbf{h}(j)\} .$$

Then, it is not hard to see that $\mathcal{C} = \bigcup_{i<j} \mathcal{C}_{i,j}$. This is because for $\mathbf{h} \in \mathcal{C}$, there exist $i < j$ such that $\mathbf{h}(i) = \mathbf{h}(j)$, and thus $\mathbf{h} \in \mathcal{C}_{i,j}$. Conversely, if $\mathbf{h} \in \mathcal{C}_{i,j}$, then $\mathbf{h}(i) = \mathbf{h}(j)$, and thus $\mathbf{h} \in \mathcal{C}$.

We are going to use the so-called *union bound*, that tells us for any two events $\mathcal{A}, \mathcal{B}$, we have $\mathbb{P}(\mathcal{A} \cup \mathcal{B}) \leqslant \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B})$. (This can be generalized to more than two events.) We can apply it here to obtain

$$\mathbb{P}(\mathcal{C}) = \mathbb{P}\left(\bigcup_{i<j} \mathcal{C}_{i,j}\right) \leqslant \sum_{i<j} \mathbb{P}(\mathcal{C}_{i,j}) . \tag{1}$$

Now, let us fix some i and j such that $1 \leqslant i < j \leqslant n$. We are going to prove that

$$\mathbb{P}(\mathcal{C}_{i,j}) = 1/M . \tag{2}$$

First observe that this is enough to conclude the proof, because we can plug this into (1) and use the fact that there are exactly $\binom{n}{2}$ pairs $i < j$ we are summing over, and thus

$$\mathbb{P}(\mathcal{C}) \leqslant \binom{n}{2}\frac{1}{M} = \frac{n(n-1)}{2M} .$$

Hence, it remains to prove (2). To this end, we need to count the number of functions $\mathbf{h} \in \mathcal{H}_{K,M}$ such that $\mathbf{h}(i) = \mathbf{h}(j)$. For this, define the event $\mathcal{A}(i, y)$ as the event that $\mathbf{h}(i) = y$. Then, note that

$$\mathbb{P}\left(\mathcal{C}_{i,j}\right) = \sum_{y \in [M]} \mathbb{P}\left(\mathcal{A}(i, y) \cap \mathcal{A}(j, y)\right) \tag{3}$$

because for $\mathcal{C}_{i,j}$ to occur, there must be some y such that $\mathbf{h}(i) = y$ and $\mathbf{h}(j) = y$. Then, also note that that there are exactly M^{K-2} functions such that $\mathbf{h}(i) = y$ and $\mathbf{h}(j) = y$, because we can freely set $\mathbf{h}(x)$ for any $x \in [K] \backslash \{i, j\}$. Thus:

$$\mathbb{P}\left(\mathcal{A}(i, y) \cap \mathcal{A}(j, y)\right) = \frac{M^{K-2}}{M^K} = \frac{1}{M^2} \, .$$

Plugging this into (3), we get $\mathbb{P}\left(\mathcal{C}_{i,j}\right) = M \cdot \frac{1}{M^2} = \frac{1}{M}$. □

Pairwise-independence. As we have seen in class, the above property is useful but it is too expensive to sample a function from the set of *all* functions, since such a function's description is large – we need to give a table of K elements from $[M]$. We would like to find sets of functions from which to sample $\mathbf{h}$ that achieve the same upper bound on the collision probability, but a function from this set can be described much more succinctly. To this end, we use the following notion.

Definition 2. We say that a set $\mathcal{H}$ of functions $[N] \to [K]$ (often called a "function family") is **pairwise-independent** if

$$\left|\left\{\mathbf{h} \in \mathcal{H} \, : \, \mathbf{h}(x) = y \ \wedge \ \mathbf{h}(x') = y'\right\}\right| = \frac{|\mathcal{H}|}{M^2} \, .$$

for all distinct x, x' in $[N]$, and all (not necessarily distinct) $y, y' \in [K]$. ◊

The point now is that if we change the above experiment to sample $\mathbf{h}$ from a pairwise-independent family $\mathcal{H}$, rather than from all functions $\mathcal{H}_{K,N}$, the above upper bound on the collision probability still holds – and the proof is very similar. This is – in abstract terms – because our proof only relies on pairwise-independent events.

Let us see why it is the case. The only place we have *really* used properties of $\mathcal{H}_{K,M}$ is when computing $\mathbb{P}\left(\mathcal{A}(i, y) \cap \mathcal{A}(j, y)\right)$. If we change the probability space so that we now have $\Omega = \mathcal{H}$, and $\mathbb{P}\left(\mathbf{h}\right) = 1/\left|\mathcal{H}\right|$ for all $\mathbf{h} \in \Omega = \mathcal{H}$, then the above definition yields

$$\mathbb{P}\left(\mathcal{A}(i, y) \cap \mathcal{A}(j, y)\right) = \frac{\left|\left\{\mathbf{h} \in \mathcal{H} \, : \, \mathbf{h}(x) = y \ \wedge \ \mathbf{h}(x') = y'\right\}\right|}{|\mathcal{H}|} = \frac{1}{M^2} \, ,$$

i.e., exactly as in the case of our proof! The point here is that we have only used the fact that the K events $\mathcal{A}(i, y)$ (for a fixed y) are pairwise-independent, and for this to be guaranteed, it is enough if we sample $\mathbf{h}$ from a pairwise-independent family $\mathcal{H}$.

But have we gained anything? One can show that $\mathcal{H}_{K,M}$ *is* pairwise-independent. (Exercise! In fact, we use this implicitly in the proof above.) But, the key point is that we can find pairwise-independent families $\mathcal{H}$ with *much smaller size*, e.g., $|\mathcal{H}| = K^2$, as opposed to M^K. In fact, theoretical constructions approaching size M^2 exist, which is obviously much smaller than M^K.