

Proof of Theorem 1

1 Appendix: Theorems

We can show that partitioning is entropy-neutral if we do not consider the order of the tuples to be important. Intuitively, splitting an unordered set of tuples into any number of smaller sets does not change the space needed to represent the tuples.

More formally, a *partitioning* of a probability distribution S with probability function p is a set of distributions induced by a partitioning of the range of S into sets S_1 through S_k , where the probability function of each partitioned distribution S_i is given by p_i . Let $p_i(s) = p(s)/P(S_i)$, where $P(S_i) = \sum_{u \in S_i} p(u)$. We denote the distribution induced on S by the partition S_i as $S|S_i$ as it acts as S conditioned on the value being in the partition S_i .

To show this we state two theorems. The first is an analogue of the Total Probability Theorem for entropy. It relates the entropy of a distribution with the entropy of partitionings of the distribution.

Theorem 1 *Given a distribution S and a partitioning of S into distributions $S|S_1$ through $S|S_k$,*

$$H(S) = \sum_{1 \leq i \leq k} P(S_i)(H(S|S_i) - \lg P(S_i)). \quad (1)$$

As in [1], we model a relation of n tuples as a multiset of n values chosen i.i.d per a probability distribution S . We use the notation $Choose(S, n)$ to denote the random variable corresponding to this multiset. A simple counting argument (e.g., see [1]) shows that:

$$\begin{aligned} H(Choose(S, n)) &= nH(S) - \lg n! \\ &\in (nH(S) - n \lg n, nH(S) - n \lg n + 2n) \end{aligned} \quad (2)$$

We now use the partitioning on S to induce a partitioning of the multiset into k separate multisets of size $N_1 \dots N_k$, one for each partition. Our goal is to show that the entropy of the original multiset $H(Choose(S, n))$ is the same as the sum of entropies of the multisets formed after partitioning $H(Choose(S|S_i, N_i))$. A tricky part of this statement is that the multiset entropies depends on the size N_i , which are themselves random variables. So we make this a theorem about expectation.

Theorem 2 *If a multiset of n values chosen i.i.d from probability distribution S is partitioned into k multisets*

of size $N_1 \dots N_k$ by a partitioning $S_1 \dots S_k$ of S , then

$$H(Choose(S, n)) + 2n \geq \sum_{i=1}^k E(H(Choose(S|S_i, N_i)))$$

where the expectation in the R.H.S is taken over all possible values of $N_1 \dots N_k$.

Proof: By (2), The LHS of the theorem $\geq nH(S) - n \lg n + 2n$.

Consider the event where N_i takes a particular value n_i . Applying (2), $H(Choose(S|S_i, n_i)) \leq n_i H(S|S_i) - n_i \lg n_i + 2n_i$ (3)

Consider the set of all compound events of the form $(N_1 \dots N_k) = (n_1 \dots n_k)$. Each occurs with probability $Prob_{(N_1 \dots N_k) = (n_1 \dots n_k)}$.

Now, the RHS of this theorem is $\sum_i E(H(Choose(S|S_i, N_i))) = Prob_{(N_1 \dots N_k) = (n_1 \dots n_k)} \sum_i H(Choose(S|S_i, n_i)) \leq \sum_i Prob_{N_i = n_i} n_i H(S|S_i) - \sum_i Prob_{N_i = n_i} (n_i \lg n_i - 2n_i)$ (by (3))

$\leq \sum_i E(N_i) H(S|S_i) + 2n - \sum_i Prob_{N_i = n_i} n_i \lg n_i$
 $\leq n \sum_i P(S_i) H(S|S_i) + 2n - \sum_i Prob(N_i = n_i) n_i \lg n_i$ (4)
 $\leq nH(S) - nH(P) + 2n - \sum_i Prob(N_i = n_i) n_i \lg n_i$ (5)

where (4) is because $E(N_i) = nP(S_i)$ and (5) follows from Theorem 1. Now, $\sum_i Prob_{N_i = n_i} n_i \lg n_i = nE(\sum_i (N_i/n) \lg N_i) = nE(\sum_i (N_i/n) (\lg N_i/n) + n \lg n \sum_i (N_i/n)) = nE(\sum_i P(S_i) (\lg 1/P(S_i)) + n \lg n \sum_i (N_i/n)) = -nH(P) + n \lg n$ (6)

Adding (5) and (6), the RHS of the theorem is $\leq nH(S) - n \lg n + 2n$. \square

References

- [1] V. Raman and G. Swart. Entropy compression of relations and querying of compressed relations. In *VLDB*, 2006.