where the units of H are in bits. A *feature* X is a discrete random variable defined on Ω. Assume that X has a finite number of values $x_1, x_2, \ldots, x_k$. Set $q_i = \mathbb{P}(X = x_i)$. The *Shannon entropy* $H(X)$ of the feature X is given by

$$H(X) = -\sum_i q_i \log_2 q_i.$$

In particular, the values of X define a partition of $\mathcal{S}$ into disjoint subsets $\mathcal{S}_i = \{s \in \mathcal{S} : X(s) = x_i\}$, for $1 \leq i \leq k$. This further induces k spaces $\Omega_i = (\mathcal{S}_i, \mathcal{P}(\mathcal{S}_i), p_i)$, where the induced distribution is given by

$$p_i(s) = \frac{p(s)}{q_i} \quad \text{for } s \in \mathcal{S}_i,$$

and q_i denotes the probability of X having value x_i and is given by

$$q_i = \mathbb{P}(X = x_i) = \sum_{s \in \mathcal{S}_i} p(s).$$

Let $H(\Omega|X)$ denote the *conditional entropy* of Ω given the value of feature X. The entropy $H(\Omega|X)$ gives the expected value of the entropies of the conditional distributions on Ω, averaged over the conditioning feature X and can be computed by

$$H(\Omega|X) = \sum_i q_i H(\Omega_i).$$

Then the *entropy reduction* $R(\Omega, X)$ of Ω for feature X is the difference between the *a priori* Shannon entropy $H(\Omega)$ and the conditional entropy $H(\Omega|X)$, i.e.

$$R(\Omega, X) = H(\Omega) - H(\Omega|X).$$

The entropy reduction indicates the change on average in information entropy from a prior state to a state that takes some information as given. Now we prove Propositions and. Given two features X_1 and X_2, we can partition Ω either first by X_1 and subsequently by X_2, or first by X_2 and then by X_1, or just by a pair of features (X_1, X_2). In the following, we will show that all three approaches provide the same entropy reduction of Ω. Before the proof, we define some notations. The joint probability distribution of a pair of features (X_1, X_2) is given by $q_{i_1,i_2} = \mathbb{P}(X_1 = x_{i_1}^{(1)}, X_2 = x_{i_2}^{(2)})$, and the marginal probability distributions are given by $q_{i_1}^{(1)} = \mathbb{P}(X_1 = x_{i_1}^{(1)})$ and $q_{i_2}^{(2)} = \mathbb{P}(X_2 = x_{i_2}^{(2)})$. Clearly, $\sum_{i_1} q_{i_1,i_2} = q_{i_2}^{(2)}$ and $\sum_{i_2} q_{i_1,i_2} = q_{i_1}^{(1)}$. The *joint entropy* $H(X_1, X_2)$ of a pair (X_1, X_2) is defined as

$$H(X_1, X_2) = -\sum_{i_1} \sum_{i_2} q_{i_1,i_2} \log_2 q_{i_1,i_2}.$$

The *conditional entropy* $H(X_2|X_1)$ of a feature X_2 given X_1 is defined as the expected value of the entropies of the conditional distributions X_2, averaged