The inequality

$$H(h_v^{(0)}{}_{\text{with features}}) \geq H(h_v^{(0)}{}_{\text{without features}})$$

can be proven as follows:

Definitions

1. **Entropy**: For a probability distribution $p(x)$, the entropy $H(p)$ is defined as:

$$H(p) = -\sum_x p(x) \log p(x),$$

where $p(x)$ is the probability of x.

2. **Initial Label Distribution**: - $h_v^{(0)}{}_{\text{without features}}$: The initial label distribution without considering node features is determined solely by structural information, such as node degree or connectivity. - $h_v^{(0)}{}_{\text{with features}}$: The initial label distribution when node features (e.g., contextual information, attributes) are included incorporates more information.

3. **Key Idea**: Including features adds more variability or information to the label distribution, potentially increasing its entropy.

Proof

1. **Entropy of Label Distribution Without Features**:

Assume $h_v^{(0)}{}_{\text{without features}}$ is based on structural information, which partitions the graph into equivalence classes of nodes with identical structure (e.g., same degree, neighbourhood, etc.). Let the label distribution over these equivalence classes be $p_{\text{struct}}(x)$.

The entropy of this distribution is:

$$H(h_v^{(0)}{}_{\text{without features}}) = -\sum_x p_{\text{struct}}(x) \log p_{\text{struct}}(x).$$

This entropy is limited by the structural variability of the graph. Nodes with identical structure will have identical labels, leading to lower entropy.

2. **Entropy of Label Distribution With Features**:

When features are added, the label distribution incorporates the variability of node features. Let this distribution be $p_{\text{features}}(x)$, which is defined over a larger space since it combines structural information and feature variability.

The entropy is:

$$H(h_v^{(0)}{}_{\text{with features}}) = -\sum_x p_{\text{features}}(x) \log p_{\text{features}}(x).$$

3. **Relationship Between $p_{\text{struct}}(x)$ and $p_{\text{features}}(x)$**:

The distribution $p_{\text{features}}(x)$ subsumes $p_{\text{struct}}(x)$ because it includes all structural information plus feature variability. This results in a more "spread-out" distribution, increasing entropy.

Mathematically:

$$p_{\text{features}}(x) = \sum_{y} p_{\text{struct}}(x \mid y) p_{\text{features}}(y),$$

where y represents feature states. The increased variability in y leads to higher entropy.

4. **Entropy Inequality**:

By the **information-theoretic property** that adding more variability (in this case, features) to a distribution increases entropy:

$$H({h_v^{(0)}}_{\text{with features}}) \geq H({h_v^{(0)}}_{\text{without features}}),$$

with equality only when features add no additional information (i.e., $p_{\text{features}}(x) = p_{\text{struct}}(x)$).

5. **Conclusion**:

Including features in $h_v^{(0)}$ increases the variability of the label distribution, leading to a more informative initialization for 1-WL updates. This increased entropy reflects a richer representation of node contexts.

Intuitive Explanation

- Without features, the label distribution depends only on structural properties, resulting in limited variability. - Adding features introduces additional sources of variability, creating a more complex and less deterministic label distribution. This increase in variability directly translates to higher entropy.