1. Linear Heuristic Function

In section 3.1 of the main paper we show that greedy search combined with the priority function \( p(v) = g(v) + h^*(v) \) will lead to the exact solution of the original ILP (Eqs. (3) to (5) in the main paper), under the condition that there is no duality gap. For convenience we repeat the definition of the optimal heuristic function \( h^*(v) \) here:

\[
h^*(v) = - \sum_{(k,i) \in v} \sum_j A_{kji}^* u_{kij}^*(x) \tag{1}
\]

In this section we show that the heuristic function in Eq. (1) can be written as a linear function of the form \(-w \cdot \phi(v)\).

First, let us recap the meanings and ranges of indices \( k, i, \) and \( j \) in Eq. (1):

- index \( k \) (from 1 to \( K \)): the \( k^{th} \) categorical variable.
- index \( i \) (from 1 to \( n \)): the \( i^{th} \) label.
- index \( j \) (from 1 to \( m \)): the \( j^{th} \) constraint.

Second, recall that a search node \( v \) is just a set of pairs \( \{(k, i)\} \), each element of which specifies that the variable \( y_k \) is assigned the \( i^{th} \) label.

Now we can define a \( K \times n \times m \) dimensional feature vector \( \phi(v) \), the component of which is labeled by a tuple of indices \( (k, i, j) \). Let the \( (k, i, j) \)th component of the feature vector be

\[
\phi_{kij}(v) = \begin{cases} 
  u_{kij}^*(x), & \text{if } (k, i) \in v \\
  0, & \text{otherwise}
\end{cases} \tag{2}
\]

Also define the corresponding weight parameter

\[
w_{kij} = A_{kji}^* \tag{3}
\]

Clearly the heuristic function in Eq. (1) is just \(-w \cdot \phi(v)\), where \( \phi \) and \( w \) is defined in Eq. (2) and Eq. (3), respectively.

2. Proof of Mistake Bound Theorem

The goal of this section is to prove Theorem 1 in the main paper. We will prove two lemmas which will lead to the final proof of the theorem. Before introducing the two lemmas, we repeat the relevant definitions here for convenience.

Let \( R_g \) be a constant such that for every pair of search nodes \( (v, v') \), \( \|\phi(v) - \phi(v')\| \leq R_g \). Let \( R_d \) be a constant such that for every pair of search nodes \( (v, v') \), \( |g(v) - g(v')| \leq R_g \).

Finally we define the level margin of a weight vector \( w \) for a training set as

\[
\gamma = \min_{\{(v, v')\}} w \cdot (\phi(v) - \phi(v')) \tag{4}
\]

Here, the set \( \{(v, v')\} \) contains any pair such that \( v \) is \( y \)-good, \( v' \) is not \( y \)-good, and \( v \) and \( v' \) are at the same search level. The priority function used to rank the search nodes is defined as \( p_w(v) = g(v) - w \cdot \phi(v) \). Smaller priority function value ranks higher during search. With these definitions we have the following two lemmas:

**Lemma 1.** Let \( w^k \) be the weights before the \( k^{th} \) mistake is made \( (w^1 = 0) \). Right after the \( k^{th} \) mistake is made, the norm of the weight vector \( w^{k+1} \) has the following upper bound:

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

**Proof.** When the \( k^{th} \) mistake is made, we get \( w^{k+1} \) by using the update rule in either line 11 or line 14 of the algorithm.

Let us consider the case of updating using line 11 first. We have

\[
\|w^{k+1}\|^2 = \|w^k + \phi(v^*) - \frac{1}{|B|} \sum_{v \in B} \phi(v)\|^2
\]

\[
= \|w^k\|^2 + \|\phi(v^*) - \frac{1}{|B|} \sum_{v \in B} \phi(v)\|^2 + 2w^k \cdot \left(\phi(v^*) - \frac{1}{|B|} \sum_{v \in B} \phi(v)\right) \tag{5}
\]

We will upper bound each term separately on the right side of Eq. (5). To bound the second term we use the definition

\[
B^{-1} = \sum_{v \in B} \phi(v) - \phi(v^*) + \phi(v^*) - \frac{1}{|B|} \sum_{v \in B} \phi(v)
\]

\[
\|B^{-1}\| \leq \|\phi(v^*)\| + \frac{1}{|B|} \sum_{v \in B} \phi(v)
\]

\[
\|\phi(v^*) - \frac{1}{|B|} \sum_{v \in B} \phi(v)\| \leq \gamma
\]

\[
\|w^k\|^2 + 2\gamma \|w^k\| + \gamma^2 \leq \|w^{k+1}\|^2
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + \gamma^2 + 2\gamma \|w^k\| + \gamma^2
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + 2\gamma \|w^k\| + \gamma^2
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]

\[
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_g^2 + 2R_g
\]
of $R_\phi$ and the properties of vector dot product:

$$
\|\phi(v^*) - \frac{1}{|B|} \sum_{v \in B} \phi(v)\|^2 \\
= \frac{1}{|B|^2} \|\sum_{v \in B} (\phi(v^*) - \phi(v))\|^2 \\
= \frac{1}{|B|^2} \sum_{v \in B} (\phi(v^*) - \phi(v)) \cdot \sum_{v \in B} (\phi(v^*) - \phi(v')) \\
= \frac{1}{|B|^2} \sum_{v, v' \in B} (\phi(v^*) - \phi(v)) \cdot (\phi(v^*) - \phi(v')) \\
\leq \frac{1}{|B|^2} \sum_{v, v' \in B} \|\phi(v^*) - \phi(v)\| \|\phi(v^*) - \phi(v')\| \\
\leq \frac{1}{|B|^2} \sum_{v, v' \in B} R_\phi^2 \\
= R_\phi^2 
$$

(6)

To bound the third term on the right side of Eq. (5), note that the update is happening because of the $k$th mistake is being made. Therefore for any node $v \in B$, we have (smaller priority function value ranks higher)

$$
p_{w^k}(v^*) \geq p_{w^k}(v)
$$

Using the definitions of $p_{w^k}(v)$ and $R_\phi$,

$$
w^k \cdot (\phi(v^*) - \phi(v)) \leq g(v^*) - g(v) \leq R_\phi
$$

Now we have the upper bound for the third term on the right side of Eq. (5):

$$
2w^k \cdot (\phi(v^*) - \frac{1}{|B|} \sum_{v \in B} \phi(v)) \\
= \frac{2}{|B|} w^k \cdot \sum_{v \in B} (\phi(v^*) - \phi(v)) \\
= \frac{2}{|B|} \sum_{v \in B} w^k \cdot (\phi(v^*) - \phi(v)) \\
\leq \frac{2}{|B|} \sum_{v \in B} R_\phi \\
= 2R_\phi
$$

(7)

Combining Eqs. (5), (6) and (7) leads to:

$$
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_\phi^2 + 2R_\phi
$$

Next we consider the case of updating using line 14 of the algorithm, in which we have

$$
\|w^{k+1}\|^2 = \|w^k + \phi(v^*) - \phi(\hat{v})\|^2 \\
= \|w^k\|^2 + \|\phi(v^*) - \phi(\hat{v})\|^2 \\
+ 2w^k \cdot (\phi(v^*) - \phi(\hat{v})) \\
= \|w^k\|^2 + R_\phi^2 + 2R_\phi
$$

(8)

Using the definition of $R_\phi$:

$$
\|\phi(v^*) - \phi(\hat{v})\|^2 \leq R_\phi^2. \\
$$

(9)

Also, since a mistake is made by the weight $w^k$, we know $p_{w^k}(v^*) \geq p_{w^k}(\hat{v})$, which implies

$$
w^k \cdot (\phi(v^*) - \phi(\hat{v})) \leq g(v^*) - g(\hat{v}) \leq R_\phi
$$

(10)

Combining Eqs. (8), (9) and (10) again leads to

$$
\|w^{k+1}\|^2 \leq \|w^k\|^2 + R_\phi^2 + 2R_\phi
$$

Lemma 2. Let $w^k$ be the weights before the $k$th mistake is made ($w^1 = 0$). Let $w$ be a weight vector with level margin $\gamma$ as defined in Eq. (4). Then

$$
w \cdot w^{k+1} \geq w \cdot w^k + \gamma
$$

Proof. First consider the update rule of line 11 of the algorithm,

$$
w \cdot w^{k+1} = w \cdot \left( w^k + \phi(v^*) - \frac{1}{|B|} \sum_{v \in B} \phi(v) \right) \\
= w \cdot w^k + \frac{1}{|B|} \sum_{v \in B} w \cdot \left( \phi(v^*) - \phi(v) \right) \\
\geq w \cdot w^k + \frac{1}{|B|} \sum_{v \in B} \gamma \\
= w \cdot w^k + \gamma
$$

where we use the definition of the level margin $\gamma$. Next consider the update rule of line 14 of the algorithm,

$$
w \cdot w^{k+1} = w \cdot \left( w^k + \phi(v^*) - \phi(\hat{v}) \right) \\
= w \cdot w^k + w \cdot \left( \phi(v^*) - \phi(\hat{v}) \right) \\
\geq w \cdot w^k + \gamma
$$

(11)
Similarly, induction on \( k \) using Lemma 2,
\[
\mathbf{w} \cdot \mathbf{w}^{k+1} \geq k \gamma.
\]
Finally we have
\[
1 \geq \frac{\mathbf{w} \cdot \mathbf{w}^{k+1}}{\|\mathbf{w}\| \|\mathbf{w}^{k+1}\|} \geq \frac{k \gamma}{\sqrt{k(R_\phi^2 + 2R_g)}}
\]
which gives us
\[
k \leq \frac{R_\phi^2 + 2R_g}{\gamma^2}
\]

3. Proof of Theorem 2 of the Main Paper

In this section we prove Theorem 2 of the main paper. We repeat the relevant definitions and the theorem here for convenience.

Given a fixed beam size \( b \) and the beam candidates \( C_t \) at step \( t \) from which we need to select the beam \( B_t \), we can rank the nodes in \( C_t \) from smallest to largest according to the heuristic function \( h(v) \). Denote the \( b \)th smallest node as \( v_b \) and the \((b + 1)\)th smallest node as \( v_{b+1} \), we define the heuristic gap \( \Delta_t \) as
\[
\Delta_t = h(v_{b+1}) - h(v_b) \tag{11}
\]
If the beam \( B_t \) is selected from \( C_t \) only according to heuristic function, then \( \Delta_t \) is the gap between the last node in the beam and the first node outside the beam. Next we define the path-cost gap \( \delta_t \) as
\[
\delta_t = \max_{v, v' \in C_t} (v - v') \tag{12}
\]
With these definitions we have the following theorem:

**Theorem 2.** Given the beam candidates \( C_t \) with heuristic gap \( \Delta_t \) and path-cost gap \( \delta_t \), if \( \Delta_t > \delta_t \), then using only heuristic function to select the beam \( B_t \) will have the same set of nodes selected as using the full priority function up to their ordering in the beam.

**Proof.** Let \( v^* \in C_t \) be any node that is ranked within top-\( b \) nodes by the heuristic function \( h(\cdot) \), and let \( v \in C_t \) be an arbitrary node that is not within top-\( b \) nodes. By definitions of \( v_b \) and \( v_{b+1} \) we have
\[
h(v^*) \leq h(v_b) \leq h(v_{b+1}) \leq h(v)
\]
Therefore,
\[
h(v) - h(v^*) \geq h(v_{b+1}) - h(v_b) = \Delta_t. \tag{13}
\]
Also by definition of \( \delta_t \) we have
\[
g(v) - g(v^*) \geq -\delta_t \tag{14}
\]
Combining Eqs. (13) and (14),
\[
p(v) - p(v^*) = \left( g(v) + h(v) \right) - \left( g(v^*) + h(v^*) \right)
\]
\[
= \left( h(v) - h(v^*) \right) + \left( g(v) - g(v^*) \right)
\]
\[
\geq \Delta_t - \delta_t
\]
Thus for any node \( v^* \in C_t \), if it is selected in the beam \( B_t \) (ranked top \( b \)) by the heuristic function \( h(\cdot) \), it will be selected in the beam \( B_t \) by the full priority function \( p(\cdot) \).