Exercise: Dual decomposition for pose segmentation

(Source: Daphne Koller.). Two important problems in computer vision are that of parsing articulated objects (*e.g.*, the human body), called *pose estimation*, and segmenting the foreground and the background, called *segmentation*. Intuitively, these two problems are linked, in that solving either one would be easier if the solution to the other were available. We consider solving these problems simultaneously using a joint model over human poses and foreground/background labels and then using dual decomposition for MAP inference in this model.

We construct a two-level model, where the high level handles pose estimation and the low level handles pixel-level background segmentation. Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected grid over the pixels. Each node $i \in \mathcal{V}$ represents a pixel. Suppose we have one binary variable x_i for each pixel, where $x_i = 1$ means that pixel i is in the foreground. Denote the full set of these variables by $\mathbf{x} = (x_i)$.

In addition, suppose we have an undirected tree structure $T = (\mathcal{V}', \mathcal{E}')$ on the parts. For each body part, we have a discrete set of candidate poses that the part can be in, where each pose is characterized by parameters specifying its position and orientation. (These candidates are generated by a procedure external to the algorithm described here.) Define y_{jk} to be a binary variable indicating whether body part $j \in \mathcal{V}'$ is in configuration k. Then the full set of part variables is given by $\mathbf{y} = (y_{jk})$, with $j \in \mathcal{V}'$ and $k = 1, \ldots, K$, where J is the total number of body parts and K is the number of candidate poses for each part. Note that in order to describe a valid configuration, $\mathbf{y}$ must satisfy the constraint that $\sum_{k=1}^{K} y_{jk} = 1$ for each j.

Suppose we have the following energy function on pixels:

$$E_1(\mathbf{x}) = \sum_{i \in \mathcal{V}} 1[x_i = 1] \cdot \theta_i + \sum_{(i,j) \in \mathcal{E}} 1[x_i \neq x_j] \cdot \theta_{ij}.$$

Assume that the θ_{ij} arises from a metric (*e.g.*, based on differences in pixel intensities), so this can be viewed as the energy for a pairwise metric MRF with respect to G.

We then have the following energy function for parts:

$$E_2(\mathbf{y}) = \sum_{p \in \mathcal{V}'} \theta_p(y_p) + \sum_{(p,q) \in \mathcal{E}'} \theta_{pq}(y_p, y_q).$$

Since each part candidate y_{jk} is assumed to come with a position and orientation, we can compute a binary mask in the image plane. The mask assigns a value to each pixel, denoted by $\{w_{jk}^i\}_{i \in \mathcal{V}}$, where $w_{jk}^i = 1$ if pixel i lies on the skeleton and decreases as we move away. We can use this to define an energy function relating the parts and the pixels:

$$E_3(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}'} \sum_{k=1}^{K} 1[x_i = 0, y_{jk} = 1] \cdot w_{jk}^i.$$

In other words, this energy term only penalizes the case where a part candidate is active but the pixel underneath is labeled as background.

Formulate the minimization of $E_1 + E_2 + E_3$ as an integer program and show how you can use dual decomposition to solve the dual of this integer program. Your solution should describe the decomposition into slaves, the method for solving each one, and the update rules for the overall algorithm. Briefly justify your design choices, particularly your choice of inference algorithms for the slaves.