

Given  $N$  tasks and  $K$  candidate groups, our method minimize the following empirical risk:

$$L^{\text{task}}(\theta, S) = L(\theta) \odot Z(S) = \sum_{i=1}^N \sum_{k=1}^K z_{ik} L_{ik}, \quad (1)$$

or the relaxed empirical risk by the Concrete distribution:

$$L^{\text{relaxed task}}(\theta, S) = L(\theta) \odot \tilde{Z}(S) = \sum_{i=1}^N \sum_{k=1}^K \tilde{z}_{ik} L_{ik}, \quad (2)$$

where  $\odot$  is the element-wise product,  $L = \{L_{ik}\} \in \mathbb{R}^{N \times K}$  is the loss matrix,  $S = \{s_{ik}\} \in \mathbb{R}^{N \times K}$  is the set of parameters of the categorical distributions,  $\theta$  is the model weights.  $Z = \{z_{ik}\} \in \mathbb{R}^{N \times K}$  and  $\tilde{Z} = \{\tilde{z}_{ik}\} \in \mathbb{R}^{N \times K}$  are the discrete and continuous relaxed sampling, respectively:

$$z_{ik} \sim \text{Categorical}(s_{ik}) \quad (3)$$

$$\tilde{z}_{ik} = \frac{\exp((s_{ik} + g_{ik})/\tau)}{\sum_{m=1}^K \exp((s_{im} + g_{im})/\tau)} \quad (4)$$

**Theorem .1.** *The convergence of Eq. (10) or (12) is no worse than the naive multi-task learning (MTL) baseline where all the tasks are categorized into one group.*

*Proof.* For the naive MTL baseline:

$$\begin{aligned} \exists k \quad \text{and} \quad \forall i, \quad z_{ik}/\tilde{z}_{ik} &= 1, \\ \text{otherwise,} \quad z_{ik}/\tilde{z}_{ik} &= 0, \end{aligned} \quad (5)$$

Equation (5) indicates that for the naive MTL baseline, in the  $N \times K$  group assignment matrix  $Z$  or  $\tilde{Z}$ , there exists a column with all **1**'s, and all **0**'s elsewhere. Given Eqs. (10) and (3) for  $Z$ , or Eqs. (12) and (4) for  $\tilde{Z}$ , it is straightforward to identify that such a circumstance is a special case of our method. By minimizing either of our empirical risks, the convergence is no worse than the naive MTL baseline.  $\square$

**Theorem .2.** *The convergence of Eq. (10) or (12) is no worse than the single-task learning (STL) baseline given  $K = N$ .*

*Proof.* For the STL baseline:

$$\begin{aligned} \forall i, \quad z_{ii}/\tilde{z}_{ii} &= 1, \\ \text{if } k \neq i, \quad z_{ik}/\tilde{z}_{ik} &= 0, \end{aligned} \quad (6)$$

Equation (6) indicates that when  $K = N$ , for the STL baseline, the diagonal of the  $N \times K$  group assignment matrix  $Z$  or  $\tilde{Z}$  are all **1**'s, and all **0**'s elsewhere. Given  $K = N$ , Eqs. (10) and (3) for  $Z$ , or Eqs. (12) and (4) for  $\tilde{Z}$ , it is straightforward to identify that such a circumstance is a special case of our method. By minimizing either of our empirical risks with  $K = N$ , the convergence is no worse than the STL baseline.  $\square$

We also note that when setting  $K = N$ , it is not necessary to categorize the input tasks into  $N$  groups in our method. Instead, our method automatically learns to categorize the input tasks into **less than**  $N$  groups that minimize the empirical risk. This is guaranteed by our categorical distribution which allows certain groups to contain **0** task, which is empirically verified in our Table 7.