A. Theorical Analysis

Given N tasks, K candidate groups, and M samples, our method minimizes the following empirical risk:

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

or the relaxed empirical risk by using the `Concrete` distribution:

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

where $\odot$ is the element-wise product, $L = \{L_{ik}^{m}\} \in \mathbb{R}^{N \times K \times M}$ is the loss tensor for each sample, each task, and each group, $S = \{s_{ik}\} \in \mathbb{R}^{N \times K}$ is the set of parameters of the `categorical` distributions, θ is the model weights. $Z = \{z_{ik}\} \in \mathbb{R}^{N \times K}$ and $\widetilde{Z} = \{\tilde{z}_{ik}\} \in \mathbb{R}^{N \times K}$ are the discrete and continuous relaxed sampling, respectively:

$$z_{ik} \sim \texttt{Categorical}(s_{ik}) \tag{3}$$

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

Theorem 1. *The convergence of Eq.* (1) *or* (2) *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:

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

Equation (5) indicates that for the naive MTL baseline, in the $N \times K$ group assignment matrix Z or $\widetilde{Z}$, there exists a column with all **1**'s, and all **0**'s elsewhere. Given Eqs. (1) and (3) for Z, or Eqs. (2) and (4) for $\widetilde{Z}$, it is straightforward to identify that such a circumstance of naive MTL is a special case of our method. By minimizing either of our empirical risks, the convergence is no worse than the naive MTL baseline. □

Theorem 2. *The convergence of Eq.* (1) *or* (2) *is no worse than the single-task learning (STL) baseline given $K = N$.*

Proof. For the STL baseline:

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

Equation (6) indicates that when $K = N$, for the STL baseline, the diagonal of the $N \times K$ group assignment matrix Z or $\widetilde{Z}$ are all **1**'s, and all **0**'s elsewhere. Given $K = N$, Eqs. (1) and (3) for Z, or Eqs. (2) and (4) for $\widetilde{Z}$, it is straightforward to identify that such a circumstance of STL 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. □

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.