Federated Multi-Task Learning

Federated Update of W

Algorithm 1 MOCHA: Federated Multi-Task Learning Framework

Input: Data $\mathbf{X}_t$ from $t = 1, \ldots, m$ tasks, stored on one of m nodes, and initial matrix $\mathbf{\Omega}_0$
Starting point $\boldsymbol{\alpha}^{(0)} := \mathbf{0} \in \mathbb{R}^n$, $\mathbf{v}^{(0)} := \mathbf{0} \in \mathbb{R}^b$
for iterations $i = 0, 1, \ldots$ **do**
    Set subproblem parameter σ' and number of federated iterations, H_i
    for iterations $h = 0, 1, \cdots, H_i$ **do**
        for tasks $t \in \{1, 2, \ldots, m\}$ **in parallel over** m **nodes do**
            call local solver, returning θ_t^h-approximate solution $\Delta\boldsymbol{\alpha}_t$ of the local subproblem (4)
            update local variables $\boldsymbol{\alpha}_t \leftarrow \boldsymbol{\alpha}_t + \Delta\boldsymbol{\alpha}_t$
            return updates $\Delta\mathbf{v}_t := \mathbf{X}_t \Delta\boldsymbol{\alpha}_t$
        reduce: $\mathbf{v}_t \leftarrow \mathbf{v}_t + \Delta\mathbf{v}_t$
    Update $\mathbf{\Omega}$ centrally based on $\mathbf{w}(\boldsymbol{\alpha})$ for latest $\boldsymbol{\alpha}$
Central node computes $\mathbf{w} = \mathbf{w}(\boldsymbol{\alpha})$ based on the lastest $\boldsymbol{\alpha}$
return: $\mathbf{W} := [\mathbf{w}_1, \ldots, \mathbf{w}_m]$

Convergence Analysis

Definition (Per-Node-Per-Iteration-Approximation Parameter)

At each iteration h, we define the accuracy level of the solution calculated by node k to its subproblem as

$$\theta_k^h := \frac{G_k^{\sigma'}(\Delta\alpha_k^{(h)}; v^{(h)}, \alpha_k^{(h)}) - G_k^{\sigma'}(\Delta\alpha_k^{\star}; v^{(h)}, \alpha_k^{(h)})}{G_k^{\sigma'}(\mathbf{0}; v^{(h)}, \alpha_k^{(h)}) - G_k^{\sigma'}(\Delta\alpha_k^{\star}; v^{(h)}, \alpha_k^{(h)})}$$

$\theta_k^h \in [0,1]$, $\theta_k^h = 1$ means that no updates to the subproblem are made at iteration h

Assumption

Let $\mathscr{H}_h := (\alpha^{(h)}, \ldots, \alpha^{(1)})$ be the *dual vector history* until the beginning of iteration h, and define $\Theta_k^h := \mathbb{E}[\theta_k^h \,|\, \mathscr{H}_h]$. For all tasks k and all iterations h, we assume $p_k^h := \mathbb{P}(\theta_k^h = 1) \leq p_{max} < 1$ and $\hat{\Theta}_k^h = \mathbb{E}[\theta_k^h \,|\, \mathscr{H}_h, \theta_k^h < 1] \leq \Theta_{max} < 1$.