Proximal Multitask Learning Over Distributed Networks with Jointly Sparse Structure

Danqi Jin[1], Jie Chen [1] , Cédric Richard [2] & Jingdong Chen [1]

[1] CIAIC, Northwestern Polytechinical University, Xi'an, China
[2] Université de la Côte d'Azur, CNRS, France

Presented by:
Danqi Jin
7th May 2020 for ICASSP 2020

danqijin@mail.nwpu.edu.cn; dr.jie.chen@ieee.org

cedric.richard@unice.fr; jingdongchen@ieee.org

Presentation Outline

Consider a connected network consisting of N nodes. Each node k has access to streaming data $\{d_{k,n}, \boldsymbol{u}_{k,n}\}$, which are related via the linear model:

$$d_{k,n} = \boldsymbol{u}_{k,n}^\top \boldsymbol{w}_k^\star + z_{k,n}. \tag{1}$$

We assume that vectors $\boldsymbol{w}_k^\star$ over the entire network are jointly sparse, namely:

$$\text{supp}(\boldsymbol{w}_1^\star) = \cdots = \text{supp}(\boldsymbol{w}_k^\star) = \cdots = \text{supp}(\boldsymbol{w}_N^\star) \tag{2}$$

where $\text{supp}(\boldsymbol{w}_k^\star) \triangleq \{j : [\boldsymbol{w}_k^\star]_j \neq 0\}$ is the support of $\boldsymbol{w}_k^\star$.

Presentation Outline

Define the local parameter matrix:

$$\boldsymbol{W}_k \triangleq \left[\boldsymbol{w}_k,\ \boldsymbol{w}_\ell^\star \text{ with } \ell \in \mathcal{N}_k^-\right] \in \mathbb{R}^{L\times|\mathcal{N}_k|}. \tag{3}$$

To facilitate the following derivation we also denote $\boldsymbol{W}_k$ by

$$\boldsymbol{W}_k = \left[\bar{\boldsymbol{w}}_{k,1}^\top \quad \cdots \quad \bar{\boldsymbol{w}}_{k,m}^\top \quad \cdots \quad \bar{\boldsymbol{w}}_{k,L}^\top\right]^\top, \tag{4}$$

where $\bar{\boldsymbol{w}}_{k,m}$ is the m-th row of matrix $\boldsymbol{W}_k$.
We consider the regularized cost at node k:

$$J_k(\boldsymbol{w}_k) = J_k'(\boldsymbol{w}_k) + \lambda_k g(\boldsymbol{w}_k) \tag{5}$$

with $J_k'(\boldsymbol{w}_k) \triangleq \frac{1}{2}\,\mathbb{E}\left\{|d_{k,n} - \boldsymbol{u}_{k,n}^\top \boldsymbol{w}_k|^2\right\}$, and $g(\boldsymbol{w}_k) \triangleq \sum_{m=1}^{L} \|\bar{\boldsymbol{w}}_{k,m}\|_\infty$ evaluates the $\ell_{\infty,1}$-norm of $\boldsymbol{W}_k$.
At each node k, we then consider the convex optimization problem:

$$\boldsymbol{w}_k^\dagger = \underset{\boldsymbol{w}_k}{\operatorname{argmin}}\, J_k(\boldsymbol{w}_k). \tag{6}$$

Proximal gradient methods generate a sequence of estimates by the following iterations:

$$\boldsymbol{w}_{k,n+1} = \operatorname{prox}_{\mu_k \lambda_k g}\big(\boldsymbol{w}_{k,n} - \mu_k \nabla J'_k(\boldsymbol{w}_{k,n})\big), \tag{7}$$

where μ_k is a positive small step-size, and the proximal operator is defined by

$$\operatorname{prox}_{\lambda_g}(\boldsymbol{v}) \triangleq \underset{\boldsymbol{w}_k}{\operatorname{argmin}}\Big(g(\boldsymbol{w}_k) + \frac{1}{2\lambda}\|\boldsymbol{w}_k - \boldsymbol{v}\|_2^2\Big). \tag{8}$$

We obtain from (7) the **proximal multitask diffusion LMS algorithm** for jointly sparse networks:

$$\begin{cases} \boldsymbol{\psi}_{k,n+1} = \boldsymbol{w}_{k,n} + \mu_k \boldsymbol{u}_{k,n}\big(d_{k,n} - \boldsymbol{u}_{k,n}^\top \boldsymbol{w}_{k,n}\big) \\ \boldsymbol{w}_{k,n+1} = \operatorname{prox}_{\mu_k \lambda_k g}\big(\boldsymbol{\psi}_{k,n+1}\big) \end{cases} \tag{9}$$

Presentation Outline

We need to derive a closed-form expression for the following proximal operator:

$$\begin{aligned} \boldsymbol{w}_{k,n+1} &= \operatorname{prox}_{\mu_k \lambda_k g}(\boldsymbol{\psi}_{k,n+1}) \\ &= \underset{\boldsymbol{w}_k}{\operatorname{argmin}} \left(g(\boldsymbol{w}_k) + \frac{1}{2\mu_k \lambda_k} \|\boldsymbol{w}_k - \boldsymbol{\psi}_{k,n+1}\|_2^2 \right). \end{aligned} \tag{10}$$

As $g(\boldsymbol{w}_k)$ is separable over its all entries, its proximal operator can be evaluated in an element-wise manner as:

$$\left[\operatorname{prox}_{\mu_k \lambda_k g}(\boldsymbol{\psi}_{k,n+1})\right]_m = \operatorname{prox}_{\mu_k \lambda_k g_m}\left([\boldsymbol{\psi}_{k,n+1}]_m\right) \tag{11}$$

with $g_m([\boldsymbol{w}_k]_m) \triangleq \|\bar{\boldsymbol{w}}_{k,m}\|_\infty$, $[\boldsymbol{w}_k]_m$ is the m-th entry of $\boldsymbol{w}_k$, and $\bar{\boldsymbol{w}}_{k,m}$ is the m-th row of matrix $\boldsymbol{W}_k$ in (3).

We have:

$$[\boldsymbol{w}_{k,n+1}]_m = \underset{[\boldsymbol{w}_k]_m}{\operatorname{argmin}} \Big(\max\{|[\boldsymbol{w}_k]_m|, |[\boldsymbol{w}^\star_\ell]_m| \text{ with } \ell \in \mathcal{N}_k^-\} + \frac{1}{2\mu_k\lambda_k}\left([\boldsymbol{w}_k]_m - [\boldsymbol{\psi}_{k,n+1}]_m\right)^2 \Big). \tag{12}$$

We denote $[\boldsymbol{w}_{k,n+1}]_m$ by $\hat{w}$ and the maximal value of $|[\boldsymbol{w}^\star_\ell]_m|$ for $\ell \in \mathcal{N}_k^-$ as $[\boldsymbol{w}^o_k]_m$.

- Case 1: $|[\boldsymbol{w}_k]_m| < [\boldsymbol{w}^o_k]_m$. In this case, (12) becomes:

$$\hat{w} = \underset{\substack{[\boldsymbol{w}_k]_m \\ |[\boldsymbol{w}_k]_m| < [\boldsymbol{w}^o_k]_m}}{\operatorname{argmin}} [\boldsymbol{w}^o_k]_m + \frac{1}{2\mu_k\lambda_k}\left([\boldsymbol{w}_k]_m - [\boldsymbol{\psi}_{k,n+1}]_m\right)^2. \tag{13}$$

The solution is directly given by:

$$\hat{w} = \begin{cases} [\boldsymbol{\psi}_{k,n+1}]_m, & \text{if } |[\boldsymbol{\psi}_{k,n+1}]_m| < [\boldsymbol{w}^o_k]_m \\ [\boldsymbol{w}^o_k]_m, & \text{if } [\boldsymbol{\psi}_{k,n+1}]_m \geq [\boldsymbol{w}^o_k]_m \\ -[\boldsymbol{w}^o_k]_m, & \text{if } [\boldsymbol{\psi}_{k,n+1}]_m \leq -[\boldsymbol{w}^o_k]_m. \end{cases} \tag{14}$$

- Case 2: $|[\boldsymbol{w}_k]_m| \geq [\boldsymbol{w}_k^o]_m$. Equation (12) becomes:

$$\hat{w} = \underset{\substack{[\boldsymbol{w}_k]_m \\ |[\boldsymbol{w}_k]_m| \geq [\boldsymbol{w}_k^o]_m}}{\operatorname{argmin}} \left(|[\boldsymbol{w}_k]_m| + \frac{1}{2\mu_k\lambda_k} \left([\boldsymbol{w}_k]_m - [\boldsymbol{\psi}_{k,n+1}]_m \right)^2 \right) \quad (15)$$

Consider first:

$$\hat{w}^o = \underset{[\boldsymbol{w}_k]_m}{\operatorname{argmin}} \left(|[\boldsymbol{w}_k]_m| + \frac{1}{2\mu_k\lambda_k} \left([\boldsymbol{w}_k]_m - [\boldsymbol{\psi}_{k,n+1}]_m \right)^2 \right) \quad (16)$$

the solution is given by the soft thresholding operator defined as:

$$\hat{w}^o = S_{\mu_k\lambda_k}\left([\boldsymbol{\psi}_{k,n+1}]_m \right) = \begin{cases} [\boldsymbol{\psi}_{k,n+1}]_m + \mu_k\lambda_k, & \text{if } [\boldsymbol{\psi}_{k,n+1}]_m < -\mu_k\lambda_k \\ [\boldsymbol{\psi}_{k,n+1}]_m - \mu_k\lambda_k, & \text{if } [\boldsymbol{\psi}_{k,n+1}]_m > \mu_k\lambda_k \\ 0 & \text{otherwise.} \end{cases} \quad (17)$$

If $[\boldsymbol{w}_k^o]_m = 0$, problem (15) becomes unconstrained and we have:

$$\hat{w} = \hat{w}^o \tag{18}$$

Otherwise, considering constraint $|[\boldsymbol{w}_k]_m| \geq [\boldsymbol{w}_k^o]_m > 0$ with (17) leads to:

$$\hat{w} = \tag{19}$$

$$\begin{cases} [\boldsymbol{\psi}_{k,n+1}]_m + \mu_k \lambda_k, & \text{if } [\boldsymbol{\psi}_{k,n+1}]_m \leq -[\boldsymbol{w}_k^o]_m - \mu_k \lambda_k \\ -[\boldsymbol{w}_k^o]_m, & \text{if } -[\boldsymbol{w}_k^o]_m - \mu_k \lambda_k < [\boldsymbol{\psi}_{k,n+1}]_m < 0 \\ -[\boldsymbol{w}_k^o]_m \text{ or } [\boldsymbol{w}_k^o]_m, & \text{if } [\boldsymbol{\psi}_{k,n+1}]_m = 0 \\ [\boldsymbol{w}_k^o]_m, & \text{if } 0 < [\boldsymbol{\psi}_{k,n+1}]_m < [\boldsymbol{w}_k^o]_m + \mu_k \lambda_k \\ [\boldsymbol{\psi}_{k,n+1}]_m - \mu_k \lambda_k, & \text{if } [\boldsymbol{\psi}_{k,n+1}]_m \geq [\boldsymbol{w}_k^o]_m + \mu_k \lambda_k \end{cases}$$

To evaluate the proximal operator (12), several issues have to be addressed.

1. We first need to know which of (14), (17) or (19) has to be applied as the proximal operator of (12).

- Case A: $[\boldsymbol{w}_k^o]_m = 0$. Since condition $|[\boldsymbol{w}_k]_m| < [\boldsymbol{w}_k^o]_m$ of Case 1 cannot hold, we only consider Case 2. The proximal operator is given by (17) directly.
- Case B: $[\boldsymbol{w}_k^o]_m > 0$. Proximal operators (14) and (19) hold simultaneously. We shall choose the solution that minimizes the cost (12). We arrive at the following expression:

$$\hat{w} = \tag{20}$$

$$\begin{cases} [\boldsymbol{\psi}_{k,n+1}]_m + \mu_k\lambda_k, & \text{if } [\boldsymbol{\psi}_{k,n+1}]_m \leq -[\boldsymbol{w}_k^o]_m - \mu_k\lambda_k \\ -[\boldsymbol{w}_k^o]_m, & \text{if } -[\boldsymbol{w}_k^o]_m - \mu_k\lambda_k < [\boldsymbol{\psi}_{k,n+1}]_m \leq -[\boldsymbol{w}_k^o]_m \\ [\boldsymbol{\psi}_{k,n+1}]_m, & \text{if } \left|[\boldsymbol{\psi}_{k,n+1}]_m\right| < [\boldsymbol{w}_k^o]_m \\ [\boldsymbol{w}_k^o]_m, & \text{if } [\boldsymbol{w}_k^o]_m \leq [\boldsymbol{\psi}_{k,n+1}]_m < [\boldsymbol{w}_k^o]_m + \mu_k\lambda_k \\ [\boldsymbol{\psi}_{k,n+1}]_m - \mu_k\lambda_k, & \text{if } [\boldsymbol{\psi}_{k,n+1}]_m \geq [\boldsymbol{w}_k^o]_m + \mu_k\lambda_k \end{cases}$$

2. Another issue is that $\hat{w}$ cannot be evaluated with (17) and (20) since $[\boldsymbol{w}_k^o]_m$ is unknown. An approximation of $[\boldsymbol{w}_k^o]_m$ is given by $\max_{\ell \in \mathcal{N}_k^-} \{ \big| [\boldsymbol{\psi}_{\ell,n+1}]_m \big| \}$.

3. Condition $[\boldsymbol{w}_k^o]_m = 0$ has to be satisfied to trigger Case A, otherwise Case B is considered. Due to the existence of gradient noise, the condition $[\boldsymbol{w}_k^o]_m = 0$ of Case A is seldom satisfied. Thus we instead use conditions $[\boldsymbol{w}_k^o]_m \leq \tau$ to trigger Case A and $[\boldsymbol{w}_k^o]_m > \tau$ to select Case B.

Presentation Outline

We considered a nonstationary jointly sparse system identification scenario with $\boldsymbol{w}_k^\star$ varying over time.

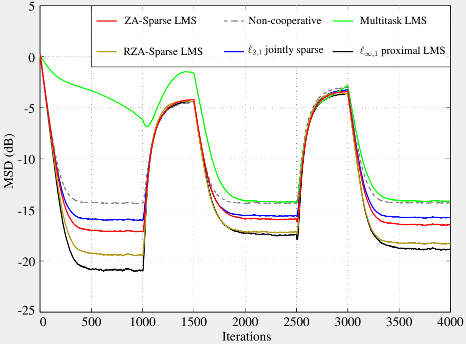

Figure 1: Simulation results with white inputs.

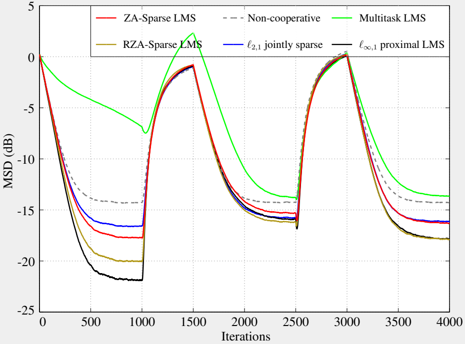

Figure 2: Simulation results with colored inputs.

Thanks for Your Times!