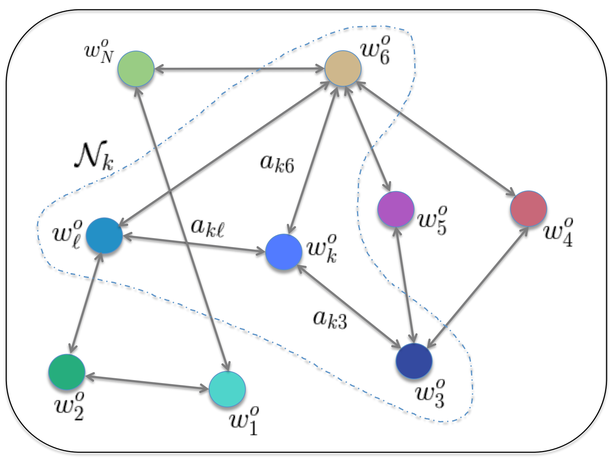
Distributed Inference over Multitask Graphs under Smoothness

Roula Nassif, Stefan Vlaski, Ali H. Sayed
Institute of Electrical Engineering, EPFL, Switzerland

Abstract

This work formulates a multitask optimization problem where agents have individual costs to minimize, subject to a smoothness condition over the graph. A diffusion strategy that responds to streaming data and employs stochastic approximations in place of actual gradient vectors is derived and studied.

Distributed Inference under Smoothness



- Connected graph $\mathcal{G} = (\mathcal{N}, A)$ with $\mathcal{N}$ a set of N nodes, and A a symmetric weighted adjacency matrix with $[A]_{k\ell} = a_{k\ell} > 0$ if there is an edge connecting k and ℓ and 0 otherwise
- Strongly convex risk $J_k(w_k) = \mathbb{E}\, Q_k(w_k; \boldsymbol{x})$ at node k with w_k^o the minimizer
- Prior belief: the signal $w^o = \text{col}\{w_1^o, \ldots, w_N^o\}$ is smooth w.r.t. the underlying weighted graph

A measure for the smoothness of w is:

$$S(w) = w^\top \mathcal{L} w = \frac{1}{2}\sum_{k=1}^{N}\sum_{\ell\in\mathcal{N}_k} a_{k\ell}\|w_k - w_\ell\|^2$$

where $\mathcal{L} = L \otimes I_M$, $L = \text{diag}\{A\mathbb{1}_N\} - A$ is the graph Laplacian, and $\mathcal{N}_k$ is the neighborhood of k.

Objective

Devise a distributed strategy that solves:

$$w_\eta^o = \arg\min_{\mathcal{W}} \sum_{k=1}^{N} J_k(w_k) + \frac{\eta}{2} w^\top \mathcal{L} w \qquad (1)$$

where $\eta \geq 0$ is a regularization strength. Agent k is interested in estimating the k-th sub-vector of $w_\eta^o = \text{col}\{w_{1,\eta}^o, \ldots, w_{N,\eta}^o\}$.

We are interested in a stochastic solution where the distribution of $\boldsymbol{x}$ is unknown, i.e., $J_k(w_k)$ and $\nabla_{w_k} J_k(w_k)$ are unknown. A common construction is to employ the following approximation at iteration i:

$$\widehat{\nabla_{w_k} J_k}(w_k) = \nabla_{w_k} Q_k(w_k; \boldsymbol{x}_i)$$

where $\boldsymbol{x}_i$ represents the data observed at iteration i.

Adaptive distributed strategy

$$\begin{cases} \boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla_{w_k} J_k}(\boldsymbol{w}_{k,i-1}) \\ \boldsymbol{w}_{k,i} = \boldsymbol{\psi}_{k,i} - \mu\eta \sum_{\ell\in\mathcal{N}_k} a_{k\ell}(\boldsymbol{\psi}_{k,i} - \boldsymbol{\psi}_{\ell,i}) \end{cases} \qquad (2)$$

where $\mu > 0$ is a small step-size.

Theoretical Motivation

Consider MSE networks where each agent is subjected to streaming data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$:

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i} w_k^o + \boldsymbol{v}_k(i)$$

with w_k^o an unknown vector and $\boldsymbol{v}_k(i)$ a measurement noise. An MSE cost is associated with node k:

$$J_k(w_k) = \frac{1}{2}\mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} w_k|^2$$

The processes $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}, \boldsymbol{v}_k(i)\}$ are assumed to be zero-mean jointly WSS satisfying: i) $\mathbb{E}\boldsymbol{u}_{k,i}^\top \boldsymbol{u}_{\ell,j} = R_{u,k}\delta_{k,\ell}\delta_{i,j}$ where $R_{u,k} > 0$; ii) $\mathbb{E}\boldsymbol{v}_k(i)\boldsymbol{v}_\ell(j) = \sigma_{v,k}^2\delta_{k,\ell}\delta_{i,j}$; iii) the regression and noise processes $\{\boldsymbol{u}_{\ell,j}, \boldsymbol{v}_k(i)\}$ are independent of each other.

Maximum a posteriori estimator

If the network parameter vector is an intrinsic Gaussian Markov Random field $\boldsymbol{w} \sim \mathcal{N}(0, \mathcal{L})$, i.e.,

$$f(\boldsymbol{w}) = (2\pi)^{-M(N-1)/2}\,(|\mathcal{L}|^*)^{1/2}\, e^{-\frac{1}{2}\boldsymbol{w}^\top\mathcal{L}\boldsymbol{w}}$$

and if the noise is Gaussian $\boldsymbol{v}_k(i) \sim \mathcal{N}(0, \sigma_{v,k}^2)$ independent over space and time and identically distributed, then problem (1) is a MAP estimator for $\boldsymbol{w}$ conditioned on $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$.

Stochastic Performance Analysis

- The risk $J_k(w_k)$ is assumed to be twice differentiable and strongly convex such that:

$$0 < \lambda_{k,\min} I_M \leq \nabla_{w_k}^2 J_k(w_k) \leq \lambda_{k,\max} I_M$$

- It is assumed that the gradient noise defined as:

$$\boldsymbol{s}_{k,i}(w_k) = \nabla_{w_k} J_k(w_k) - \widehat{\nabla_{w_k} J_k}(w_k)$$

satisfies:

$$\mathbb{E}[\boldsymbol{s}_{k,i}(\boldsymbol{w}_k)|\boldsymbol{\mathcal{F}}_{i-1}] = 0$$

$$\mathbb{E}[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_k)\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \leq \beta_k^2\|\boldsymbol{w}_k\|^2 + \sigma_{s,k}^2$$

for some $\beta_k^2 \geq 0$, $\sigma_{s,k}^2 \geq 0$, and where $\boldsymbol{\mathcal{F}}_{i-1}$ denotes the filtration generated by the random processes $\{\boldsymbol{w}_{\ell,j}\}$ for all $\ell = 1, \ldots, N$ and $j \leq i-1$

Stochastic performance

Strategy (2) induces a contraction mapping when:

$$0 \leq \mu\eta \leq \frac{2}{\lambda_{\max}(L)}, \text{ and } 0 < \mu < \min_{1\leq k\leq N}\left\{\frac{2}{\lambda_{k,\max}}\right\}$$

and leads to small estimation errors:

$$\text{MSD} \triangleq \limsup_{i\to\infty} \frac{1}{N}\mathbb{E}\|w_\eta^o - \boldsymbol{w}_i\|^2 = O(\mu)$$

Simulation Results

Let $\boldsymbol{\gamma}_k = \pm 1$ denote a class random variable and $\boldsymbol{h}_k$ denote the corresponding feature vector. During the training phase, k receives $\{\boldsymbol{\gamma}_k(i), \boldsymbol{h}_{k,i}\}$ with $\boldsymbol{h}_{k,i} = \boldsymbol{\gamma}_k(i)\cdot r\cdot \text{col}\{\cos(\theta_k), \sin(\theta_k)\} + \boldsymbol{v}_{k,i}$, $\boldsymbol{v}_{k,i}$ is drawn from $\mathcal{N}(0, \sigma_{v,k}^2 I)$ and $\boldsymbol{\gamma}_k(i)$ is Bernoulli distributed with $p(\boldsymbol{\gamma}_k(i) = +1) = 0.5$. We set $r = \sqrt{2}$ and $\theta_k = \frac{\pi}{6} + \frac{k-1}{N-1}\cdot\frac{7\pi}{6}$. We are interested in finding a decision rule, parameterized by w_k^o, such that $\hat{\boldsymbol{\gamma}}_k(i) = \text{sign}(\boldsymbol{h}_{k,i}^\top w_k^o)$ and

$$w_k^o \triangleq \arg\min_{w_k} \mathbb{E}\ln\left(1 + e^{-\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}^\top w_k}\right) + \rho\|w_k\|^2.$$

We consider a network of 50 nodes where k is connected to $k-1$ and $k+1$ if $k \neq \{1, 50\}$, node 1 is connected to 2, and node 50 is connected to 49. The weight over a link is set to 1/3. We set $\mu = 10^{-3}$ and $\rho = 0.025$.

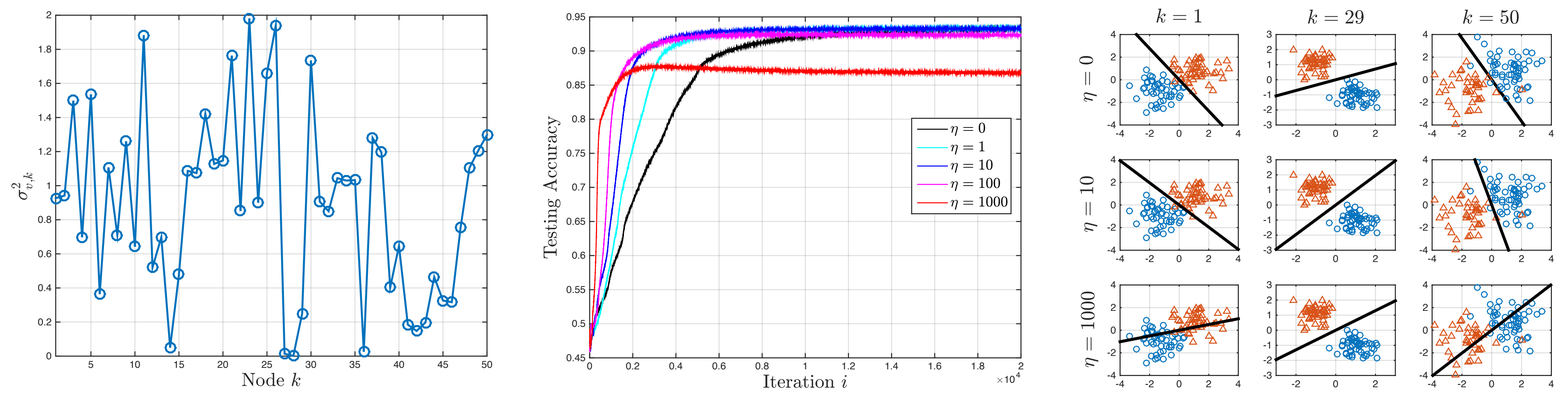


Figure 1: *(Left)* Noise profile. *(Middle)* Classification accuracy. *(Right)* One realization (after convergence) of the classifier. Blue and red circles correspond to feature vectors of 100 test samples at node k.

Remarks

- Long versions of this work have been submitted and can be found on arXiv [2, 3]. The results in Part II reveal explicitly the influence of the network topology, the data characteristics, and the regularization strength on the network performance and provide insights into the design of multitask strategies.
- A connection with graph signal processing is provided in the paper and in the longer version [3].

References

[1] A. H. Sayed.
Adaptation, learning, and optimization over networks.
Foundations and Trends in Machine Learning, 7(4-5):311–801, 2014.

[2] R. Nassif, S. Vlaski, and A. H. Sayed.
Learning over multitask graphs – Part I: Stability analysis.
Submitted for publication. Available as arXiv:1805.08535, May 2018.

[3] R. Nassif, S. Vlaski, and A. H. Sayed.
Learning over multitask graphs – Part II: Performance analysis.
Submitted for publication. Available as arXiv:1805.08547, May 2018.