MULTITASK LEARNING WITH LEARNED TASK RELATIONSHIPS

Zirui Wan and Stefan Vlaski

Department of Electrical and Electronic Engineering, Imperial College London, UK

ABSTRACT

Classical consensus-based strategies for federated and decentralized learning are statistically suboptimal in the presence of heterogeneous local data or task distributions. As a result, in recent years, there has been growing interest in multitask or personalized strategies, which allow individual agents to benefit from one another in pursuing locally optimal models without enforcing consensus. Existing strategies require either precise prior knowledge of the underlying task relationships or are fully non-parametric and instead rely on meta-learning or proximal constructions. In this work, we introduce an algorithmic framework that strikes a balance between these extremes. By modeling task relationships through a Gaussian Markov Random Field with an unknown precision matrix, we develop a strategy that jointly learns both the task relationships and the local models, allowing agents to self-organize in a way consistent with their individual data distributions. Our theoretical analysis quantifies the quality of the learned relationship, and our numerical experiments demonstrate its practical effectiveness.

Index Terms— Decentralized multitask learning, graph signal processing, Gaussian Markov Random Field.

1. INTRODUCTION

Classical federated learning and decentralized learning formulations often impose a consensus constraint—i.e., all agents must agree on a single global model or decision [1–3]. However, under heterogeneous local data or diverse task distributions, strict consensus can be statistically suboptimal because it enforces an overly restrictive compromise across tasks [4,5]. This limitation has motivated *multitask learning*, in which each agent maintains its own parameter vector while exploiting inter-task relationships through structural priors [4–7]. By explicitly modeling these relationships and incorporating them into the learning process, multitask learning can achieve superior estimation accuracy compared to consensus-based approaches.

In this work, we consider networked learning problems where each agent aims to estimate a parameter vector by minimizing its own individual cost:

$$w_k^o = \arg\min_{w_k} \ J_k(w_k). \tag{1}$$

When prior knowledge is available, it can be incorporated into multitask learning through a regularization term that augments the objective with a penalty enforcing desirable structure in the solution [6,8]. In particular, we consider:

$$\arg\min_{\mathcal{W}} \ \mathcal{J}(\mathcal{w}) \ + \ \frac{\eta}{2} \mathcal{w}^\top \mathcal{R} \, \mathcal{w}, \tag{2}$$

where $\mathcal{w} = \mathrm{col}\{w_1, \ldots, w_K\}$ concatenates $w_k \in \mathbb{R}^M$ from K agents into a column vector, $\mathcal{J}(\mathcal{w}) \triangleq \sum_{k=1}^{K} J_k(w_k)$ is the aggregate cost, $\eta > 0$ is a tuning parameter, and $\mathcal{R}$ is a positive semidefinite matrix. In this work we adopt a *graph smoothness* regularizer by setting $\mathcal{R}$ as a graph Laplacian matrix. This choice promotes similarity of the parameters across neighboring agents while still allowing heterogeneity.

Optimization with graph Laplacian regularization has been extensively studied. For example, [6–9] analyze iterative solutions and characterize how topology, stepsize, and regularization strength affect the steady-state performance. A key limitation of these works is their assumption of full knowledge of the Laplacian (both connectivity and edge weights), which is often unavailable in practice. Other efforts attempt to learn the Laplacian directly, typically from structured data over the graph [10,11]. In contrast, our work departs from these settings: the estimated Laplacian is not tied to data relationships but instead encodes latent *task relationships* among agents, and must be inferred from noisy, non-cooperative parameter estimates, resulting in a coupled problem of inferring both task relationships and optimal parameters.

Another line of work adopts non-parametric strategies based on meta-learning [12–14] or proximal formulations [15,16]. Meta-learning methods adapt models across tasks by searching a common launch model. Proximal approaches, on the other hand, control personalization through penalties on the deviation from a reference model. While flexible and able to accommodate complex task structures, these methods typically require larger datasets and provide limited interpretability of the underlying task relationships.

Motivated by these challenges, we propose a strategy that jointly learns local models and their inter-task relationships. The dependencies among tasks are modeled through a Gaussian Markov Random Field (GMRF) whose *unknown* precision matrix is constrained to the space of valid graph Laplacians and inferred from non-cooperative estimates of the local models. The estimated Laplacian is subsequently incorporated into the decentralized multitask learning procedure to promote structured cooperation among agents. We establish bounds on the Laplacian estimation error in the small-stepsize and high-dimensional regimes, showing an $O(\mu)$ dependence on the non-cooperative learning stepsize μ. Finally, we evaluate the downstream learning performance when using the estimated Laplacian and compare it against several baseline methods. The proposed framework has potential applications in large-scale sensor networks, recommendation systems, and federated healthcare analytics [17–19], where agents must learn related but non-identical models while exploiting latent structural relationships.

Throughout the paper, all vectors are column vectors. Random quantities are in boldface; matrices are uppercase, and vectors/scalars are lowercase. $\otimes$ denotes the Kronecker product, $\mathrm{diag}(\cdot)$ constructs a block-diagonal matrix, and $\mathrm{vec}(\cdot)$ stacks the columns of a matrix into a vector. The notation $\| \cdot \|$ refers to the spectral norm for matrices and the ℓ_2-norm for vectors.

2. GAUSSIAN MARKOV RANDOM FIELD

Assume that the true graph parameter $w^o = \text{col}\{w_1^o, \ldots, w_K^o\}$ is modeled by a Gaussian Markov Random Field (GMRF). In this framework, each agent's parameter w_k is treated as a Gaussian random variable, with conditional dependence relations encoded by the edges of a connected, undirected graph. This construction captures the intuition that neighboring agents are more likely to have similar parameters, thereby promoting smoothness across the network.

Each edge (k, ℓ) on the graph is assigned a nonnegative weight $a_{k\ell}$ reflecting the similarity between agents k and ℓ. Let A denote the weighted adjacency matrix with entries $A_{k\ell} = a_{k\ell}$, and $D = \text{diag}(d_1, \ldots, d_K)$ with $d_k = \sum_\ell a_{k\ell}$. The graph Laplacian is then defined as

$$L \triangleq D - A. \tag{3}$$

The Laplacian is adopted as the precision matrix of the GMRF, directly tying the probabilistic model to the graph topology [10,20,21]. Accordingly, we state the following assumption:

Assumption 1 (GMRF model). *The true parameter vector $\boldsymbol{w}^o$ is assumed to follow a multivariate Gaussian distribution:*

$$\boldsymbol{w}^o \sim \mathcal{N}\{0, \mathcal{L}^\dagger\}, \tag{4}$$

where $\mathcal{L} = L \otimes I_M$ and $(\cdot)^\dagger$ denotes the pseudo-inverse. Its probability density function is given by

$$f(w^o) = \frac{\exp\left(-\frac{1}{2}(w^o)^\top \mathcal{L} w^o\right)}{\sqrt{\det^*(2\pi \mathcal{L}^\dagger)}}, \tag{5}$$

where $\det^*(\cdot)$ denotes the pseudo-determinant and δ is the mean vector.

Since the Kronecker structure $\mathcal{L} = L \otimes I_M$ replicates the same graph-induced dependency across all feature dimensions, distribution (4) also implies that every feature dimension provides an *independent sample* of the dependency structure encoded by the graph Laplacian. As a result, the empirical covariance across features concentrates around $L^\dagger$, and the estimation error decays at the classical sub-Gaussian rate $O(K/M)$ [22].

3. MULTITASK LEARNING ALGORITHM

Under the GMRF prior (4), the maximum a posteriori (MAP) estimate of w^o naturally takes the form of (2). The MAP estimator is given by [6,23]

$$w_i^* = \underset{\mathcal{W}^o}{\arg\min} \left\{ -\log f_{\{x_j\}_{j=1}^i | \mathcal{W}^o}(\{x_j\}_{j=1}^i \mid w^o) - \log f(w^o) \right\} \tag{6}$$

$$= \underset{\mathcal{W}}{\arg\min} \ \mathcal{Q}(w; \{x_j\}_{j=1}^i) + \tfrac{1}{2} w^\top \mathcal{L} w, \tag{7}$$

here, $\{x_j\}_{j=1}^i$ denotes the collection of data observed by all agents up to time i. By substituting the prior (4) into (6) and defining the instantaneous loss as $\mathcal{Q}(w; \{x_j\}_{j=1}^i) \triangleq -\log f_{\{x_j\}_{j=1}^i | \mathcal{W}^o}(\{x_j\}_{j=1}^i \mid w^o)$, we obtain the regularized multitask formulation (7).

The cost function is then defined as the expected loss $\mathcal{J}(w) \triangleq \mathbb{E}_{x_j} \mathcal{Q}(w; x_j)$. This leads to the Laplacian-regularized optimization problem:

$$w^* = \underset{\mathcal{W}}{\arg\min} \ \mathcal{J}(w) + \frac{1}{2} w^\top \mathcal{L} w, \tag{8}$$

which matches the formulation in (2).

We are particularly interested in solving (8) in the stochastic setting, where the data distribution—and hence the exact cost and gradient $\nabla \mathcal{J}(w)$—are unknown. In this case, agents implement a stochastic gradient descent recursion [6,7]:

$$\boldsymbol{w}_i = (I_{MK} - \mu\mathcal{L})\boldsymbol{w}_{i-1} - \mu\widehat{\nabla\mathcal{J}}(\boldsymbol{w}_{i-1}), \tag{9}$$

where $\mu > 0$ is the stepsize and $\widehat{\nabla\mathcal{J}}(w)$ is a stochastic approximation of the gradient based on the available data. Due to the sparse structure of the graph Laplacian, the resulting algorithm is *decentralized* by design. When $\mathcal{L}$ is known, this algorithm converges (for sufficiently small μ) to the optimal MAP solution [6].

4. LAPLACIAN ESTIMATION

However, in practice, we cannot assume full knowledge of the Laplacian L. Instead, we rely on the structural prior in (4) to estimate a suitable $\widehat{L}$. Since the true parameter vectors w^o that encode inter-task relationships are not directly accessible, they must first be estimated locally in a non-cooperative manner, without knowledge of L. This is accomplished through a stochastic gradient descent recursion performed independently at each agent:

$$\boldsymbol{w}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \widehat{\nabla J_k}(\boldsymbol{w}_{k,i-1}), \tag{10}$$

where each agent updates its own parameter estimate using only local data. The resulting estimates are then aggregated to approximate the covariance $\Sigma = L^\dagger$. Under the zero-mean model, we adopt the empirical covariance estimator:

$$\widehat{\Sigma} = \frac{1}{M} \ \text{blktr}\big((P \, w_i)(P \, w_i)^\top\big), \tag{11}$$

where $P \in \mathbb{R}^{KM \times MK}$ is the commutation matrix that reshapes w_i into an element-wise stacking, and $\text{blktr}(\cdot)$ denotes the block-trace operator that sums the $K \times K$ diagonal blocks. According to (3), L is symmetric and positive semidefinite with $\text{rank}(L) = K - 1$ and $\text{Null}(L) = \text{Null}(L^\dagger) = \text{span}\{\mathbb{1}\}$. In contrast, the empirical covariance estimate is full rank since additive noise fills the null space. To mitigate the noise amplification in the pseudo-inverse caused by rank mismatch [24], a subspace projection should be applied to $\widehat{\Sigma}$:

$$\widehat{\Sigma}^\perp \triangleq Q \, \widehat{\Sigma} \, Q, \ Q \triangleq I_K - \frac{1}{K}\mathbb{1}\mathbb{1}^\top. \tag{12}$$

Finally, the Laplacian estimate is obtained as

$$\widehat{L} \triangleq \big(\widehat{\Sigma}^\perp\big)^\dagger, \ \widehat{\mathcal{L}} = \widehat{L} \otimes I_M. \tag{13}$$

The learned Laplacian is then incorporated into the decentralized multitask recursion:

$$\widehat{\boldsymbol{w}}_i = (I_{MK} - \mu\widehat{\mathcal{L}})\widehat{\boldsymbol{w}}_{i-1} - \mu\widehat{\nabla\mathcal{J}}(\widehat{\boldsymbol{w}}_{i-1}), \tag{14}$$

5. ESTIMATION QUALITY ANALYSIS

Since (10) produces noisy parameter estimates, we examine how this noise propagates into the covariance and Laplacian estimation by measuring the mean-squared-errors $\mathbb{E}_{\boldsymbol{\mathcal{W}}_i}\|\widehat{\Sigma}^\perp - L^\dagger\|^2$ and $\mathbb{E}_{\boldsymbol{\mathcal{W}}_i}\|\widehat{L} - L\|$. These expectations characterize the effect of stochastic fluctuations in $\boldsymbol{w}_i$. To establish this analysis, we introduce the following assumptions on the cost $J_k(\cdot)$ and the gradient noise process $\boldsymbol{s}_i(\cdot)$, defined as

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) = \widehat{\nabla\mathcal{J}}(\boldsymbol{w}_{i-1}) - \nabla\mathcal{J}(\boldsymbol{w}_{i-1}). \tag{15}$$

These assumptions are commonly satisfied by objective functions of interest in learning and adaptation, such as quadratic and logistic costs [3].

Assumption 2 (Cost functions)**.** *Each individual cost $J_k(w_k)$ is assumed to be convex, twice differentiable, and with bounded Hessian satisfying:*

$$\nu I_M \leq \nabla^2 J_k(w_k) \leq \delta I_M, \tag{16}$$

where $0 < \nu \leq \delta < \infty$.

2) The Hessian $\nabla^2 J_k(\cdot)$ satisfies the Lipschitz condition for any $w_1, w_2 \in \mathbb{R}^M$ and $k_H \geq 0$:

$$\|\nabla^2 J_k(w_1) - \nabla^2 J_k(w_2)\| \leq k_H \|w_1 - w_2\|. \tag{17}$$

Assumption 3 (Gradient noise)**.** *For any $\boldsymbol{w}_{i-1}$, gradient noise satisfies:*

$$\mathbb{E}[\boldsymbol{s}_i(\boldsymbol{w}_{i-1})|\boldsymbol{w}_{i-1}] = 0 \tag{18}$$

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4|\boldsymbol{w}_{i-1}] \leq \beta^4 \|w^o - \boldsymbol{w}_{i-1}\| + \sigma_s^4, \tag{19}$$

where $\beta, \sigma_s \geq 0$.

2) The conditional covariance of $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$ defined as

$$\mathcal{R}_{s,i}(\boldsymbol{w}_{i-1}) \triangleq \mathbb{E}[\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\boldsymbol{s}_i^\top(\boldsymbol{w}_{i-1})|\mathbb{F}_{i-1}]$$

satisfies:

$$\|\mathcal{R}_{s,i}(\boldsymbol{w}_{i-1}) - \mathcal{R}_{s,i}(w^o)\| \leq k_s \|\boldsymbol{w}_{i-1} - w^o\|^{\gamma_s} \tag{20}$$

$$\mathcal{R}_s \triangleq \lim_{i \to \infty} \mathcal{R}_{s,i}(w^o) > 0, \tag{21}$$

where $k_s \geq 0$ and $0 < \gamma_s \leq 4$.

Under the Assumption 2 and 3, we can call on the following Theorem from [25–27].

Theorem 1 (Asymptotic Normality)**.** *For sufficiently small stepsize μ and as $i \to \infty$, the sequence generated by (10) converges in distribution to an approximately conditional Gaussian [25, 26]:*

$$\boldsymbol{w}_i \mid w^o \sim \mathcal{N}(w^o, \Pi), \tag{22}$$

where Π denotes the steady-state error covariance matrix, which depends on the realization of the true parameter w^o. In particular, Π is the unique symmetric positive semidefinite solution to the discrete Lyapunov equation [27]:

$$U\Pi U - \Pi + \mu^2 \mathcal{R}_s = 0, \tag{23}$$

$$U \triangleq I_{KM} - \mu\mathcal{H}, \quad \mathcal{H} = \text{diag}(\nabla^2 J_k(w_k^o)). \tag{24}$$

Moreover, Π vanishes linearly with the stepsize [27]:

$$\Pi = O(\mu) \tag{25}$$

Theorem 1 shows that the parameter estimates become asymptotically Gaussian, regardless of the distribution of the underlying data. This Gaussian approximation enables a tractable analysis of the subsequent error propagation.

Lemma 1 (Covariance estimation error)**.** *Suppose Assumption 1 through 3 hold. For $M \gg K$, the projected sample covariance estimator in (11) satisfies:*

$$\mathbb{E}_{\boldsymbol{W}_i}\|\widehat{\Sigma}^\perp - L^\dagger\|^2 \leq \mathbb{E}_{\boldsymbol{W}^o}\Big[\underbrace{c_1\big(\|\Phi\|^2 + \|L^\dagger\|^2\big)\big(\frac{K}{M} + \frac{K^2}{M^2}\big)}_{\text{covariance concentration}}$$

$$+ \underbrace{c_2\big(\text{tr}(L^\dagger)\,\text{tr}(\Phi) + \|\frac{1}{M}\,\text{blktr}(\Phi)\|^2\big)}_{\text{bias}}\Big], \tag{26}$$

where $c_1, c_2 \geq 0$ are nonnegative constants, and $\Phi \triangleq P\Pi P^\top$, with P denoting the commutation matrix.

Proof. Omitted due to space limitations. □

Since Φ is a permutation of the steady-state error covariance Π, it also depends on the true parameter realization w^o. The additional expectation $\mathbb{E}_{\boldsymbol{W}^o}[\cdot]$ on the right-hand side of (26) accounts for this dependency, ensuring that the bound in Lemma 1 holds uniformly over both the non-cooperative estimates $\boldsymbol{w}_i$ and ture parameters w^o.

The error bound in Lemma 1 offers useful insights into the quality of covariance estimation. In particular, the estimator $\widehat{\Sigma}^\perp$ is not consistent for finite stepsizes μ: even as the number of samples $M \to \infty$, certain terms on the right-hand side remain non-vanishing, leaving a nonzero *bias* in the limit. More precisely, the first term in (26) decays at rate $O\big(\frac{K}{M}\big)$ due to covariance concentration [22], while the last term persists and contributes to the asymptotic bias:

$$\lim_{M \to \infty} \mathbb{E}_{\boldsymbol{W}_i}\|\widehat{\Sigma}^\perp - L^\dagger\|^2 = c_2\big(\text{tr}(L^\dagger)\mathbb{E}_{\boldsymbol{W}^o}[\text{tr}(\Phi)] + \mathbb{E}_{\boldsymbol{W}^o}\|\frac{1}{M}\,\text{blktr}(\Phi)\|^2\big). \tag{27}$$

Since Π vanishes with the stepsize as in (25), it follows that $\Phi = O(\mu)$. Consequently, the bias term decreases with the stepsize at rate $O(\mu)$, and thus:

$$\lim_{\mu \to 0} \lim_{M \to \infty} \mathbb{E}_{\boldsymbol{W}_i}\|\widehat{\Sigma}^\perp - L^\dagger\|^2 = 0. \tag{28}$$

Theorem 2 (Laplacian estimation error)**.** *Suppose the conditions of Lemma 1 hold. By combining the covariance error bound with the pseudo-inverse perturbation results in Theorem 4.1 of [24], we obtain that, for sufficiently small stepsize μ and sufficiently large sample size M:*

$$\mathbb{E}_{\boldsymbol{W}_i}\|\widehat{L} - L\|^2 \overset{[24]}{\leq} c_3 \|L\|^2 \,\mathbb{E}_{\boldsymbol{W}_i}\|(\widehat{\Sigma}^\perp)^\dagger\|^2 \,\mathbb{E}_{\boldsymbol{W}_i}\|\widehat{\Sigma}^\perp - L^\dagger\|^2 \tag{29}$$

$$= O(\mu). \tag{30}$$

In particular, analogous to (28), the asymptotic bias vanishes in the joint limit:

$$\lim_{\mu \to 0} \lim_{M \to \infty} \mathbb{E}_{\boldsymbol{W}_i}\|\widehat{L} - L\|^2 = 0. \tag{31}$$

Proof. Omitted due to space limitations. □

From the bounds in (26)–(31), we conclude that with sufficiently large M and sufficiently small stepsize μ, the Laplacian estimation error decays at order $O(\mu)$ and vanishes in the joint limit $\mu \to 0$, $M \to \infty$. At the same time, the dependence on $\text{tr}(L^\dagger)$ and $\|L\|^2$ highlights the role of graph structure. $\text{tr}(L^\dagger)$ is the sum of the inverses of the nonzero Laplacian eigenvalues. It becomes large when the graph is weakly connected, since small nonzero eigenvalues inflate the sum. In contrast, $\|L\|^2$ is controlled by the largest node degree and edge weights, and therefore increases in graphs with high-degree hubs or heavily weighted edges. Thus, graphs with poor connectivity or highly unbalanced structure amplify estimation errors, making them intrinsically more difficult to recover accurately.

6. SIMULATION RESULTS

We construct a connected, undirected network with $K = 10$ agents and maximum node degree 8 shown in Figure 1. To create an unbalanced weighted topology that better reflects heterogeneous task relationships and provides a more challenging test for the multitask learning algorithm, each edge weight is drawn from a mixture of two
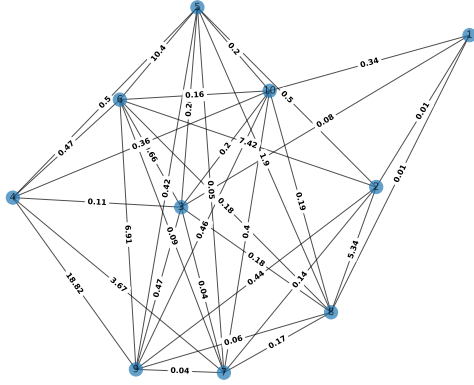
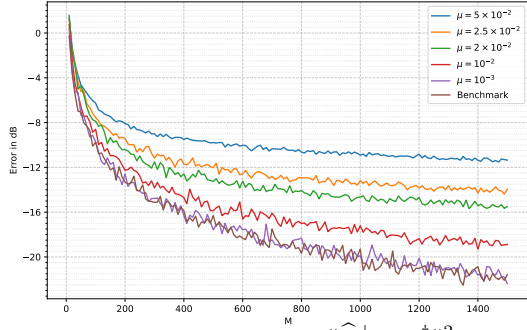**Fig. 1**: Graph topology with assigned edge weights.

Fig. 2: Covariance estimation error $\|\widehat{\Sigma}^{\perp} - L^{\dagger}\|^2$ versus M.

Fig. 3: Laplacian estimation error $\|\widehat{L} - L\|^2$ versus M.

Fig. 4: Learning performance of different algorithms.

uniform distributions: with probability 0.3 from $\text{unif}(1, 20)$ (large weights), and otherwise from $\text{unif}(0, 0.5)$ (small weights).

Each agent is subjected to streaming data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$, satisfying a linear regression model:

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i}^{\top} w_k^o + \boldsymbol{v}_k(i), \ k = 1, ..., K. \qquad (32)$$

The processes $\{\boldsymbol{u}_{k,i}, \boldsymbol{v}_k(i)\}$ are zero-mean jointly wide-sense stationary with: i) $\mathbb{E}\boldsymbol{u}_{k,i}\boldsymbol{u}_{\ell,i}^{\top} = \sigma_{u,k}^2 I_M$ if $k = \ell$ and zero otherwise; ii) $\mathbb{E}\boldsymbol{v}_k(i)\boldsymbol{v}_\ell(i) = \sigma_{v,k}^2$ if $k = \ell$ and zero otherwise; iii) $\boldsymbol{u}_{k,i}$ and $\boldsymbol{v}_k(i)$ are independent. According to (6), the cost functions take the mean-square-error form:

$$J_k(w_k) = \frac{1}{2}\mathbb{E}|d_k(i) - \boldsymbol{u}_{k,i}^{\top} w_k|^2. \qquad (33)$$

We run algorithm (10) with different stepsizes $\mu \in \{5 \times 10^{-2}, 2.5 \times 10^{-2}, 2 \times 10^{-2}, 10^{-2}, 10^{-3}\}$ until convergence, and the resulting estimates $\boldsymbol{w}_i$ are used in (11) and (12) to approximate the true covariance Σ. These estimates are compared against a benchmark constructed from the true $\boldsymbol{w}^o$. Figure 2 provides a numerical verification of the bound in (26). The y-axis reports the covariance estimation error $\|\widehat{\Sigma}^{\perp} - L^{\dagger}\|^2$, averaged over 100 Monte-Carlo trials of $\boldsymbol{w}_i$ and 100 realizations of $\boldsymbol{w}^o$ drawn from the GMRF model (4) to approximate the expectation. We observe that when M is sufficiently large, the estimation error converges to the bias terms characterized in (27), scaling consistently with the theoretical trend $O(\mu)$. This trend is evident in Figure 2; for instance, at $M = 1500$, halving the stepsize μ reduces the error magnitude by a factor of $1/2$, which corresponds to a decrease of approximately -3 dB in the error curve. Conversely, when μ is very small and M is not sufficiently large, the error is
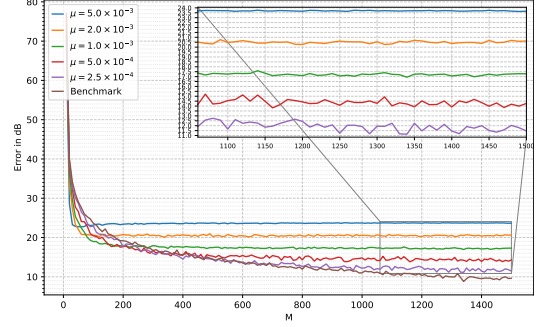
dominated by the covariance concentration terms, which scale as $O\left(\frac{K}{M}\right)$. This explains why the curve corresponding to $\mu = 10^{-3}$ nearly overlaps with the benchmark.

Figure 3 presents a numerical verification of the bound in (30). Laplacian estimation results with $\mu \in \{5 \times 10^{-3}, 2 \times 10^{-3}, 10^{-3}, 5 \times 10^{-4}, 2.5 \times 10^{-4}\}$ are compared against the benchmark. Similar to Figure 2, the Laplacian estimation errors $\|\widehat{L} - L\|^2$ are averaged over $\boldsymbol{w}_i$ and $\boldsymbol{w}^o$. We observe that when the stepsize μ is sufficiently small, the steady-state error also follows the theoretical trend $O(\mu)$.

Finally, Figure 4 shows the transient learning performance, measured by the mean-squared deviation $\frac{1}{K}\mathbb{E}\|\boldsymbol{w}^o - \boldsymbol{w}_i\|^2$, for different algorithms under the same adaptation stepsize 2×10^{-2} and $M = 1500$. The comparison includes the non-cooperative recursion (10), the consensus strategy [3], and the multitask recursion (9) with Laplacians of varying estimation accuracy. The results show that the proposed multitask strategy generally learns faster by leveraging the estimated Laplacian to coordinate updates among statistically related agents, and that higher-quality Laplacian estimates yield correspondingly better performance.

7. CONCLUSION

In this work, we proposed a distributed multitask learning framework that jointly estimates local models and the underlying task relationships. By modeling inter-task dependencies through a GMRF model, we derived a practical strategy that learns the graph Laplacian from non-cooperative estimates. Our theoretical analysis established bounds on the Laplacian estimation error, highlighting their dependence on the stepsize and feature dimensions. Simulation results validated these bounds and further demonstrated that the learned Laplacian enables faster and more effective adaptation compared to non-cooperative and consensus strategies.

8. REFERENCES

[1] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[2] P. Braca, S. Marano, and V. Matta, "Enforcing consensus while monitoring the environment in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3375–3380, 2008.

[3] A. H. Sayed et al., "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[4] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4129–4144, 2014.

[5] J. Plata-Chaves, A. Bertrand, M. Moonen, S. Theodoridis, and A. M. Zoubir, "Heterogeneous and multitask wireless sensor networks—Algorithms, applications, and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 3, pp. 450–465, 2017.

[6] R. Nassif, S. Vlaski, C. Richard, and A. H. Sayed, "Learning over multitask graphs—Part I: Stability analysis," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 28–45, 2020.

[7] R. Nassif, S. Vlaski, C. Richard, and A. H. Sayed, "Learning over multitask graphs—Part II: Performance analysis," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 46–63, 2020.

[8] R. Nassif, S. Vlaski, C. Richard, and A. H Sayed, "A regularization framework for learning over multitask graphs," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 297–301, 2018.

[9] A. J. Smola and R. Kondor, "Kernels and regularization on graphs," in *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop*. Springer, 2003, pp. 144–158.

[10] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning laplacian matrix in smooth graph signal representations," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6160–6173, 2016.

[11] D. I. Shuman, S. K. Narang, P. Frossard, et al., "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[12] M. Kayaalp, S. Vlaski, and A. H. Sayed, "Dif-MAML: Decentralized multi-agent meta-learning," *IEEE Open Journal of Signal Processing*, vol. 3, pp. 71–93, 2022.

[13] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. NeurIPS*, 2020, vol. 33, pp. 3557–3568.

[14] M. Khodak, M. F. Balcan, and A. S. Talwalkar, "Adaptive gradient-based meta-learning methods," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] C. T-Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with moreau envelopes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21394–21405, 2020.

[16] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[17] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks-Part I: Gaussian case," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 1131–1143, 2006.

[18] S. Wang, L. Hu, Y. Wang, et al., "Graph learning based recommender systems: A review," *arXiv preprint arXiv:2105.06339*, 2021.

[19] M. J. Sheller, G. A. Reina, B. Edwards, et al., "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 92–104.

[20] H. Rue and L. Held, *Gaussian Markov random fields: theory and applications*, Chapman and Hall/CRC, 2005.

[21] C. Zhang and D. Florêncio, "Analyzing the optimality of predictive transform coding using graph-based models," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 106–109, 2012.

[22] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge University Press, 2018.

[23] R. Nassif, S. Vlaski, and A. H. Sayed, "Distributed inference over multitask graphs under smoothness," in *Proc. IEEE International Workshop on Signal Processing Advances in Wireless Communications*. IEEE, 2018, pp. 1–5.

[24] P. Wedin, "Perturbation theory for pseudo-inverses," *BIT Numerical Mathematics*, vol. 13, no. 2, pp. 217–232, 1973.

[25] A. N. Shiryaev, *Probability*, vol. 95, Springer, 2016.

[26] H. J. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, Springer, 2003.

[27] X. Zhao and A. H. Sayed, "Distributed clustering and learning over networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 13, pp. 3285–3300, 2015.