Experimental Design Networks: A Paradigm for Serving Heterogeneous Learners under Networking Constraints

Yuanyuan Li*, *Member, IEEE,* Yuezhou Liu*, *Member, IEEE,* Lili Su, *Member, IEEE,* Edmund Yeh, *Senior Member, IEEE,* and Stratis Ioannidis, *Member, IEEE,*

Abstract—Significant advances in edge computing capabilities enable learning to occur at geographically diverse locations. In general, the training data needed in those learning tasks are not only heterogeneous but also not fully generated locally. In this paper, we propose an *experimental design network* paradigm, wherein learner nodes train possibly different *Bayesian linear regression models* via consuming data streams generated by data source nodes over a network. We formulate this problem as a social welfare optimization problem in which the global objective is defined as the sum of experimental design objectives of individual learners, and the decision variables are the data transmission strategies subject to network constraints. We first show that, assuming *Poisson* data streams *in steady state*, the global objective is a continuous DR-submodular function. We then propose a Frank-Wolfe type algorithm that outputs a solution within a $1 - 1/e$ factor from the optimal. Our algorithm contains a novel gradient estimation component which is carefully designed based on Poisson tail bounds and sampling. Finally, we complement our theoretical findings through extensive experiments. Our numerical evaluation shows that the proposed algorithm outperforms several baseline algorithms both in maximizing the global objective and in the quality of the trained models.

Index Terms—Experimental Design, DR-submodularity, Bayesian linear regression.

1 Introduction

We study a network in which heterogeneous learners dispersed at different locations perform local learning tasks by fetching relevant yet remote data. Concretely, data sources generate data streams containing both features and labels/responses, which are transmitted over the network (potentially through several intermediate router nodes) towards learner nodes. Generated data samples are used by learners to train models locally. We are interested in the design of rate allocation strategies that maximize the model training quality of learner nodes, subject to network constraints. This problem is relevant in practice. For example, in a mobile edge computing network [2], [3], data are generated by end devices such as mobile phones (data sources) and sent to edge servers (learners) for model training, a relatively intensive computation. In a smart city [4], [5], we can collect various types of data such as image, temperature, humidity, traffic, and seismic measurements, from different sensors. These data could be used to forecast transportation traffic, the spread of disease, pollution levels, the weather, and so on, while training for each task could happen at different public service entities.

We quantify the impact that data samples have on learner model training accuracy by leveraging objectives motivated by *experimental design* [6], a classic problem in statistics and machine learning. This problem arises in many machine learning and data mining settings, including recommender systems [7], active learning [8], and data preparation [9], to name a few. In standard experimental design, a learner decides on which experiments to conduct so that, under budget constraints, an objective modeling prediction accuracy is maximized. Learner objectives are usually scalarizations of the estimation error covariance.

In this paper, we propose *experimental design networks*, a novel optimization framework that extends classic experimental problems to maximize the sum of experimental design objectives across networked learners. Assuming Poisson data streams and Bayesian linear regression as the learning task, we define the utility of a learner as the expectation of its so-called D-optimal design objective [6], namely, the log-determinant of the learner's estimation error covariance matrix. Our goal is to determine the data rate allocation of each network edge that maximizes the aggregate utility across learners. Extending experimental design for networked learners is non-trivial. Literature on experimental design for machine learning considers budgets imposed on the number of data samples used to train the model [10]–[14]. Instead, we consider far more complex constraints on the data transmission rates across the network, as determined by network link capacities, the network topology, and data generation rates at sources.

To the best of our knowledge, we are the first to study such a networked learning problem, wherein learning tasks at heterogeneous learners are coupled via data transmission constraints over an arbitrary network topology. Our detailed contributions are as follows:

- We are the first to introduce and formalize the experimental design network problem, which enables the

• *The authors are with the Electrical and Computer Engineering Department, Northeastern University, Boston, MA 02115 USA (e-mail: yuanyuanli@ece.neu.edu; liu.yuez@northeastern.edu; l.su@northeastern.edu; eyeh@ece.neu.edu; ioannidis@ece.neu.edu).*

** Y. Li and Y. Liu contributed equally to the paper.*

This is an extended version of a paper that appeared in the IEEE International Conference on Computer Communications (INFOCOM 2022) [1].

study of multi-hop data transmission strategies for distributed learning over arbitrary network topologies.
- We prove that, assuming *Poisson* data streams *in steady state*, *Bayesian linear regression* as the learning task, and D-optimal design objectives at the learners, our framework leads to the maximization of continuous DR-submodular objective subject to a lower-bounded convex constraint set.
- Though the objective is not concave, we propose a polynomial-time algorithm based on a variant of the Frank-Wolfe algorithm [15]. To do so, we introduce and analyze a novel gradient estimation procedure, tailored to Poisson data streams. We show that the proposed algorithm, coupled with our novel gradient estimation, is guaranteed to produce a solution within a $1 - 1/e$ approximation factor from the optimal.
- We conduct extensive evaluations over different network topologies, showing that our proposed algorithm outperforms several baselines both in maximizing the objective function and in the quality of trained target models.

The rest of this paper is organized as follows. In Sections 2 and 3, we review related work and provide technical preliminaries. Section 4 introduces our framework of experimental design networks. Section 5 describes our proposed algorithm. We discuss extensions in Section 6 and present numerical experiments in Section 7. We conclude in Section 8.

2 Related Work

Distributed Computing/Learning in Networks. Distribution of computation tasks has been studied in hierarchical edge cloud networks [16], multi-cell mobile networks [17], and joint with data caching in arbitrary networks [18]. There is a rich literature on distributing machine learning computations over networks, including exchanging gradients in federated learning [19]–[21], states in reinforcement learning [22], and data vs. model offloading [23] among collaborating neighbor nodes. We depart from the aforementioned works in (a) considering multiple learners with distinct learning tasks, (b) introducing experimental design objectives, quite different from objectives considered above, (c) studying a multi-hop network, and (d) focusing on the optimization of streaming data movements, as opposed to gradients or intermediate result computations.

Experimental Design. The experimental design problem is classic and well-studied [6], [24]. Several works study the D-optimality objective [10]–[13], [25] for a single learner subject to budget constraints on the cost for conducting the experiments. Departing from previous work, we study a problem involving multiple learners subject to more complex constraints, induced by the network. Our problem also falls in the *continuous* DR-submodular setting, departing from the discrete setting in prior work. In fact, our work is the first to show that such an optimization, with Poisson data streams, can be solved via continuous DR-submodularity techniques.

DR-submodular Optimization. Submodularity is traditionally studied in the context of set functions [26], [27], but was recently extended to functions over the integer lattice [28] and the continuous domain [15]. Despite the non-convexity and the general NP-hardness of the problem, when the constraint set is down-closed and convex, maximizing monotone continuous DR-submodular functions can be done in polynomial time via a variant of the Frank-Wolfe algorithm. This yields a solution within $1 - 1/e$ from the optimal [15], [27], outperforming the projected gradient ascent method, which provides $1/2$ approximation guarantee over arbitrary convex constraints [29].

The continuous greedy algorithm [27] maximizes a submodular set function subject to matroid constraints: this first applies the aforementioned Frank-Wolfe variant to the so-called *multilinear relaxation* of the discrete submodular function, and subsequently uses rounding [30], [31]. The multilinear relaxation of a submodular function is in fact a canonical example of a continuous DR-submodular function, whose optimization comes with the aforementioned guarantees. Our objective function results from a *new continuous relaxation*, which we introduce in this paper for the first time. In particular, we show that assuming a Poisson distribution on inputs on the (integer lattice) DR-submodular function of D-optimal design yields a continuous DR-submodular function. This "Poisson" relaxation is directly motivated by our networking problem, is distinct from the multilinear relaxation [27], [29], [32], and requires a novel gradient estimation procedure. Our constraint set also requires special treatment as it is not down-closed, as required by the aforementioned Frank-Wolfe variant [15], [27]; nevertheless, we attain a $1 - 1/e$ approximation, improving upon the $1/2$ factor of projected gradient ascent [29].

Submodularity in Networking and Learning. Submodular functions are widely encountered in studies of both networking and machine learning. Submodular objectives appear in studies of network caching [33], [34], routing [35], rate allocation [36], sensor network design [37], as well as placement of virtual network functions [38]. Submodular utilities are used for data collection in sensor networks [39] and also the design of incentive mechanisms for mobile phone sensing [40]. Many machine learning problems are submodular, including structure learning, clustering, feature selection, and active learning (see e.g., [41]). Our proposed experimental design network paradigm expands this list in a novel way.

3 Technical Preliminary

We begin with a technical preliminary on linear regression, experimental design, and DR-submodularity. The contents of this section are classic; for additional details, we refer the interested reader to, e.g., [42], [43] for linear regression, [6] for experimental design, and [15] for submodularity.

3.1 Bayesian Linear Regression

In the standard linear regression setting, a learner observes n samples $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ are the feature vector and label of sample i, respectively. Labels are assumed to be linearly related to the features; in particular, there exists a model parameter vector $\boldsymbol{\beta} \in \mathbb{R}^d$ such that

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \text{for all } i \in \{1, \ldots, n\}, \tag{1}$$

and ϵ_i are i.i.d. zero mean normal noise variables with variance $\sigma^2 \in \mathbb{R}_+$ (i.e., $\epsilon_i \sim N(0, \sigma^2)$).

The learner's goal is to estimate the model parameter $\boldsymbol{\beta}$ from samples $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$. In Bayesian linear regression, it is additionally assumed that $\boldsymbol{\beta}$ is sampled from a prior normal distribution with mean $\boldsymbol{\beta}_0 \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{d\times d}$ (i.e., $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$). Under this prior, given dataset $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$, maximum a posteriori (MAP) estimation of $\boldsymbol{\beta}$ amounts to [43]:

$$\hat{\boldsymbol{\beta}}_{\texttt{MAP}} = (\boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{X}^\top \boldsymbol{y} + \sigma^2 \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0), \quad (2)$$

where $\boldsymbol{X} = [\boldsymbol{x}_i^\top]_{i=1}^n \in \mathbb{R}^{n\times d}$ is the matrix of features, $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of labels, σ^2 is the noise variance, and $\boldsymbol{\beta}_0$, $\boldsymbol{\Sigma}_0$ are the mean and covariance of the prior, respectively. We note that, in practice, the inherent noise variance σ^2 is often not known, and is typically treated as a regularization parameter and determined via cross-validation.

The quality of this estimator can be characterized by the covariance of the estimation error difference $\hat{\boldsymbol{\beta}}_{\texttt{MAP}} - \boldsymbol{\beta}$ (see, e.g., Eq. (10.55) in [43]):

$$\texttt{cov}(\hat{\boldsymbol{\beta}}_{\texttt{MAP}} - \boldsymbol{\beta}) = \big(\frac{1}{\sigma^2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\big)^{-1} \in \mathbb{R}^{d\times d}. \quad (3)$$

The covariance summarizes estimator quality in all directions in $\mathbb{R}^d$: given an unseen sample $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \mathbb{R}$, also obeying (1), the expected prediction error (EPE) is given by

$$\mathbb{E}\left[(y - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_{\texttt{MAP}})^2\right] = \sigma^2 + \boldsymbol{x}^\top \texttt{cov}(\hat{\boldsymbol{\beta}}_{\texttt{MAP}} - \boldsymbol{\beta})\boldsymbol{x}. \quad (4)$$

Hence, the eigenvalues of Eq. (3) capture the overall variability of the expected prediction error in different directions.

3.2 Experimental Design

In experimental design, a learner with prior $N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ on parameters $\boldsymbol{\beta}$ determines which experiments to conduct to learn the most accurate linear model. Formally, given p possible experiment settings, each described by feature vectors $\boldsymbol{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, p$, the learner selects a total of n experiments to conduct with these feature vectors, possibly with repetitions,[1] collects associated labels, and then performs linear regression on these sample pairs. In classic experimental design (see, e.g., [6]), the selection is formulated as an optimization problem minimizing a scalarization of the covariance, given by Eq. (3). For example, in D-optimal design, the vector $\boldsymbol{n} = [n_i]_{i=1}^p \in \mathbb{N}^p$ of the number of times each experiment is to be performed is determined by minimizing

$$\log\det[\texttt{cov}(\hat{\boldsymbol{\beta}}_{\texttt{MAP}} - \boldsymbol{\beta})] \overset{(3)}{=} \log\det\left[\big(\sum_{i=1}^p \frac{n_i}{\sigma^2}\boldsymbol{x}_i\boldsymbol{x}_i^\top + \boldsymbol{\Sigma}_0^{-1}\big)^{-1}\right]$$

or, equivalently, by solving the maximization problem:

$$\text{Max.: } G(\boldsymbol{n}; \sigma, \boldsymbol{\Sigma}_0) \equiv \log\det\big(\textstyle\sum_{i=1}^p \frac{n_i}{\sigma^2}\boldsymbol{x}_i\boldsymbol{x}_i^\top + \boldsymbol{\Sigma}_0^{-1}\big), \quad (5a)$$

$$\text{s.t.: } \textstyle\sum_{i=1}^p n_i = n. \quad (5b)$$

This equivalence follows from the fact that $\det(\boldsymbol{A}^{-1}) = (\det(\boldsymbol{A}))^{-1}$ and, thus, $\log\det(\boldsymbol{A}^{-1}) = -\log\det(\boldsymbol{A})$.

1. Note that, due to the presence of noise in labels, repeating the same experiment makes intuitive sense; formally, repetition of an experiment with features $\boldsymbol{x}_i$ reduces the EPE (4) in this direction.

Interpreting the D-Optimality Objective. In summary, Problem (5) selects $\boldsymbol{n} \in \mathbb{N}^p$ in a way so that the $\log\det[\texttt{cov}(\hat{\boldsymbol{\beta}}_{\texttt{MAP}} - \boldsymbol{\beta})]$ is as small as possible. In effect this amounts to selecting the experiments that *minimize the product of the eigenvalues of the covariance of the MAP estimator, given by Eq.* (3).[2] As discussed in Sec. 3.1, the covariance characterizes the quality of the linear estimator, as it summarizes the expected prediction error in all possible directions. Therefore, selecting experiments this way leads to the most accurate MAP linear estimate.

There are however alternative interpretations of this objective. Prob. (5) also maximizes the mutual information between the labels $\boldsymbol{y}$ (to be collected) and $\hat{\boldsymbol{\beta}}_{\texttt{MAP}}$ [10]: in this interpretation too, the selection aims to pick experiments in a way that minimizes the variability of the resulting estimator $\hat{\boldsymbol{\beta}}_{\texttt{MAP}}$. Finally, the objective itself can be interpreted as a form of coverage in the space $\mathbb{R}^d$: it aims to select experiments $\boldsymbol{X}$ that span directions not already covered by $\boldsymbol{\Sigma}_0^{-1}$: these are precisely the directions at which the prior is most uncertain.

3.3 DR-Submodularity

We introduce here diminishing-returns submodularity:

Definition 1 (DR-Submodularity [15], [28])**.** *A function $f : \mathbb{N}^p \to \mathbb{R}$ is called* diminishing-returns (DR) submodular *iff for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{N}^p$ such that $\boldsymbol{x} \leq \boldsymbol{y}$ and all $k \in \mathbb{N}$,*

$$f(\boldsymbol{x}{+}k\boldsymbol{e}_j){-}f(\boldsymbol{x}) \geq f(\boldsymbol{y}{+}k\boldsymbol{e}_j){-}f(\boldsymbol{y}), \text{ for all } j = 1, \ldots, p, \quad (6)$$

where $\boldsymbol{e}_j$ is the j-th standard basis vector.

Moreover, if (6) *holds for a real valued function $f : \mathbb{R}_+^p \to \mathbb{R}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ such that $\boldsymbol{x} \leq \boldsymbol{y}$ and all $k \in \mathbb{R}_+$, the function is called* continuous DR-submodular*.*

The above definition generalizes the submodularity of set functions (whose domain is $\{0, 1\}^p$) to functions over integer lattice (in the case of DR-submodularity), and continuous functions (in the case of continuous DR-submodularity). Particularly for continuous functions, if f is differentiable, continuous DR-submodularity is equivalent to ∇f being antitone. Moreover, if f is twice-differentialble, f is continuous DR-submodular if all entries of its Hessian $\nabla^2 f$ are non-positive. DR-submodularity is directly pertinent to D-optimal design:

Lemma 1 (Horel et al. [10])**.** *Function $G : \mathbb{N}^p \to \mathbb{R}_+$ in* (5a) *is (a) monotone-increasing and (b) DR-submodular.*

For completeness, we provide a proof in Appendix A. Problem (5) is a classic NP-hard problem [10]. An immediate consequence of this lemma is that polynomial-time approximation algorithms exist to solve Prob. (5) with a $1 - 1/e$ guarantee (see, e.g., [10], [32]), although Prob. (5) is a classic NP-hard problem [10].

4 Problem Formulation

We consider a network that aims to facilitate a distributed learning task. The network comprises (a) data source nodes

2. Other commonly encountered scalarizations [6] behave similarly. E.g., E-optimality minimizes the maximum eigenvalue, while A-optimality minimizes the sum of the eigenvalues.

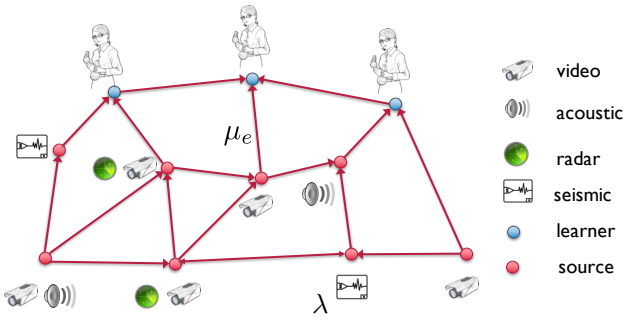


Fig. 1. An example of a experimental design network. Red nodes are sources and blue nodes are learners. Sources are generated streams of labeled data from diverse sensors, e.g., camera, radar, microphone, etc., with rate λ. Learners train their own models consuming these data. Each link e has its link capacity μ_e. The experimental design network is to allocate the data traffic for learning model better.

(e.g., sensors, test sites, experimental facilities, etc.) that generate streams of data, (b) learner nodes, that consume data with the purpose of training models, and (c) intermediate nodes (e.g., routers), that facilitate the communication of data from sources to learners. The data that learners wish to consume is determined by experimental design objectives, akin to the ones described in Sec. 3.2. Our goal is to design network communications in an optimal fashion, that maximizes learner social welfare (i.e., the sum of utilities across learners). We describe each of the above system components in more detail below.

4.1 Network Model.

We model the above system as a general multi-hop network, shown in Fig. 1, with a topology represented by a directed acyclic graph (DAG) $\mathcal{G}(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}\subset\mathcal{V}\times\mathcal{V}$ is the set of links. Each link $e=(u,v)\in\mathcal{E}$ can transmit data at a maximum rate (i.e., link capacity) $\mu^e\geq 0$. Sources $\mathcal{S}\subset\mathcal{V}$ of the DAG (i.e., nodes with no incoming edges) generate data streams, while learners $\mathcal{L}\subset\mathcal{V}$ reside at DAG sinks (nodes with no outgoing edges). We assume this for simplicity; we discuss how to remove this assumption, and how to generalize our analysis beyond DAGs, in Sec. 6.

Data Sources. Each data source $s\in\mathcal{S}$ generates a stream of labeled data. In particular, we assume that there exists a finite set $\mathcal{X}\subset\mathbb{R}^d$ of experiments every source can conduct. Once experiment with features $\boldsymbol{x}\in\mathcal{X}$ is conducted, the source can label it with a label $y\in\mathbb{R}$ of type t out of a set of possible types $\mathcal{T}$. Intuitively, features $\boldsymbol{x}$ correspond to parameters set in an experiment (e.g., pixel values in an image, etc.), label types $t\in\mathcal{T}$ correspond to possible measurements (e.g., temperature, radiation level, etc.), and labels y correspond to the actual measurement value collected (e.g., 23°C).

We assume that every source generates labeled pairs $(\boldsymbol{x},y)\in\mathbb{R}^d\times\mathbb{R}$ of type t according to a Poisson process of rate $\lambda^s_{\boldsymbol{x},t}\geq 0$. Moreover, we assume that generated data follows a linear model (1); that is, for every type $t\in\mathcal{T}$, there exists a $\boldsymbol{\beta}_t\in\mathbb{R}^d$ s.t. $y=\boldsymbol{x}^\top\boldsymbol{\beta}_t+\epsilon_t$ where $\epsilon_t\in\mathbb{R}$ are i.i.d. zero mean normal noise variables with variance $\sigma_t^2>0$, independent across experiments and sources $s\in\mathcal{S}$.

We assume $\mathcal{X}$ is finite both for simplicity, but also to highlight the connection of our networked setting to the classic experimental design problem presented in Section 3.2. This case is also of practical interest, as it holds when, e.g., features are categorical or one-hot encoded/binarized, or when continuous features are quantized. Nevertheless, our analysis also applies to (even uncountably) infinite $\mathcal{X}$: we extend our model to a setting in Sec. 6.

Learners. Each learner $\ell\in\mathcal{L}$ wishes to learn a model $\boldsymbol{\beta}_{t^\ell}$ for some type $t^\ell\in\mathcal{T}$. We assume that each learner has a prior $N(\boldsymbol{\beta}_\ell,\boldsymbol{\Sigma}_\ell)$ on $\boldsymbol{\beta}_{t^\ell}$. The learner wishes to use the network to receive data pairs $(\boldsymbol{x},y)$ of type t^ℓ, and subsequently estimate $\boldsymbol{\beta}_{t^\ell}$ through the MAP estimator (2). Note that two learners ℓ,ℓ' may be interested to learn the same model (if $t^\ell=t^{\ell'}$).

Network Constraints. The different data pairs $(\boldsymbol{x},y)\in\mathbb{R}^d\times\mathbb{R}$ generated by sources are transmitted over edges in the network along with their types $t\in\mathcal{T}$ and eventually delivered to learners. Our network design aims at allocating network capacity to different flows to meet learner needs.[3] For each edge $e\in\mathcal{E}$, we denote the rate with which data pairs of type $t\in\mathcal{T}$ with features $\boldsymbol{x}\in\mathcal{X}$ are transmitted as $\lambda^e_{\boldsymbol{x},t}\geq 0$. We also denote by

$$\lambda^{v,\mathtt{in}}_{\boldsymbol{x},t}\equiv\begin{cases}\lambda^v_{\boldsymbol{x},t}, & \text{if } v\in\mathcal{S},\\ \sum_{(u,v)\in\mathcal{E}}\lambda^{(u,v)}_{\boldsymbol{x},t}, & \text{o.w.}\end{cases}\tag{7}$$

the corresponding incoming traffic to node $v\in\mathcal{V}$, and

$$\lambda^{v,\mathtt{out}}_{\boldsymbol{x},t}=\sum_{(v,u)\in\mathcal{E}}\lambda^{(v,u)}_{\boldsymbol{x},t}\tag{8}$$

the corresponding outgoing traffic from $v\in\mathcal{V}$. Specifically for learners, we denote by

$$\lambda^\ell_{\boldsymbol{x}}\equiv\lambda^{\ell,\mathtt{in}}_{\boldsymbol{x},t^\ell},\text{ and }\boldsymbol{\lambda}^\ell=[\lambda^\ell_{\boldsymbol{x}}]_{\boldsymbol{x}\in\mathcal{X}}\in\mathbb{R}_+^{|\mathcal{X}|},\quad\text{for all }\ell\in\mathcal{L},\tag{9}$$

the incoming traffic with different features $\boldsymbol{x}\in\mathcal{X}$ of type t^ℓ at $\ell\in\mathcal{L}$. To satisfy capacity constraints, we must have

$$\sum_{\boldsymbol{x}\in\mathcal{X},t\in\mathcal{T}}\lambda^e_{\boldsymbol{x},t}\leq\mu^e,\quad\text{for all }e\in\mathcal{E},\tag{10}$$

while flow bounds imply that

$$\lambda^{v,\mathtt{out}}_{\boldsymbol{x},t}\leq\lambda^{v,\mathtt{in}}_{\boldsymbol{x},t},\quad\text{for all }\boldsymbol{x}\in\mathcal{X},t\in\mathcal{T},v\in\mathcal{V}\setminus\mathcal{L},\tag{11}$$

as data pairs can be dropped. We denote by

$$\boldsymbol{\lambda}=\left[[\lambda^e_{\boldsymbol{x},t}]_{\boldsymbol{x}\in\mathcal{X},t\in\mathcal{T},e\in\mathcal{E}};[\lambda^\ell_{\boldsymbol{x}}]_{\boldsymbol{x}\in\mathcal{X},\ell\in\mathcal{L}}\right]\tag{12}$$

the vector comprising edge and learner rates. Let

$$\mathcal{D}=\left\{\boldsymbol{\lambda}\in\mathbb{R}_+^{|\mathcal{X}||\mathcal{T}||\mathcal{E}|}\times\mathbb{R}_+^{|\mathcal{X}||\mathcal{L}|}\text{ that satisfy (9)–(11)}\right\},\tag{13}$$

be the feasible set of edge rates and learner rates. We make following assumption on the network substrate:

Assumption 1. *For $\boldsymbol{\lambda}\in\mathcal{D}$, the system is stable and, in steady state, pairs $(\boldsymbol{x},y)\in\mathbb{R}^d\times\mathbb{R}$ of type t^ℓ arrive at learner $\ell\in\mathcal{L}$ according to $|\mathcal{X}|$ independent Poisson processes with rate $\lambda^\ell_{\boldsymbol{x}}$.*

This is satisfied, if, e.g., the network is a Kelly network [44] of $M/M/1$ queues, $M/M/c$ queues, etc., under FIFO, Last-In First-Out (LIFO), and processor sharing service disciplines, or other queues for which Burke's theorem holds [43].

3. We assume hop-by-hop routing; see Sec. 6 for an extension to source routing.

TABLE 1
Notation Summary

$\mathcal{G}(\mathcal{V},\mathcal{E})$	Network graph with nodes $\mathcal{V}$ and edges $\mathcal{E}$
μ^e	Link capacity of link e
$(\boldsymbol{x}, y)$	Labeled pairs with features $\boldsymbol{x}$ and label y
s, $\mathcal{S}$	Data source index and set of all data sources
ℓ, $\mathcal{L}$	Learner index and set of all learner nodes
t, $\mathcal{T}$	Label type index and set of all types
$\lambda^e_{\boldsymbol{x},t}$	Rate of data pairs of type t with features $\boldsymbol{x}$ over edge e
$\lambda^s_{\boldsymbol{x},t}$	Rate of data pairs of type t with features $\boldsymbol{x}$ generated by source s
$\lambda^\ell_{\boldsymbol{x}}$	Rate of data pairs of type t^ℓ with features $\boldsymbol{x}$ received by learner ℓ
$n^\ell_{\boldsymbol{x}}$	The cumulative number of times that a pair $(\boldsymbol{x}, y)$ of type t^ℓ collected by learner ℓ
U^ℓ	Utility of learner ℓ

4.2 Networked Learning Problem

We consider a data acquisition time period T, at the end of which each learner $\ell \in \mathcal{L}$ estimates $\boldsymbol{\beta}_{t^\ell}$ based on the data it has received during this period via MAP estimation. Under Assumption 1, the arrivals of pertinent data pairs at learner ℓ form a Poisson process with rate $\lambda^\ell_{\boldsymbol{x}}$. Let $n^\ell_{\boldsymbol{x}} \in \mathbb{N}$ be the cumulative number of times that a pair $(\boldsymbol{x}, y)$ of type t^ℓ was collected by learner ℓ during this period, and $\boldsymbol{n}^\ell = [n^\ell_{\boldsymbol{x}}]_{\boldsymbol{x}\in\mathcal{X}}$ the vector of arrivals across all experiments. Then,

$$\Pr[\boldsymbol{n}^\ell = \boldsymbol{n}] = \prod_{\boldsymbol{x}\in\mathcal{X}} \frac{(\lambda^\ell_{\boldsymbol{x}} T)^{n_{\boldsymbol{x}}} e^{-\lambda^\ell_{\boldsymbol{x}} T}}{n_{\boldsymbol{x}}!}, \tag{14}$$

for all $\boldsymbol{n} = [n_{\boldsymbol{x}}]_{\boldsymbol{x}\in\mathcal{X}} \in \mathbb{N}^{|\mathcal{X}|}$ and $\ell \in \mathcal{L}$. Motivated by standard experimental design (see Sec. 3.2), we define the utility at learner $\ell \in \mathcal{L}$ as the following expectation:

$$\begin{aligned} U^\ell(\boldsymbol{\lambda}^\ell) &= \mathbb{E}_{\lambda^\ell}\left[G^\ell(\boldsymbol{n}^\ell)\right] \\ &= \sum_{\boldsymbol{n}\in\mathbb{N}^{|\mathcal{X}|}} G^\ell(\boldsymbol{n}) \cdot \Pr[\boldsymbol{n}^\ell = \boldsymbol{n}], \end{aligned} \tag{15}$$

where $G^\ell(\boldsymbol{n}^\ell) \equiv G(\boldsymbol{n}^\ell; \sigma_{t^\ell}, \boldsymbol{\Sigma}_\ell)$ and G is given by (5a). We wish to solve the following problem:

$$\text{Maximize:} \quad U(\boldsymbol{\lambda}) = \sum_{\ell\in\mathcal{L}} (U^\ell(\boldsymbol{\lambda}^\ell) - U^\ell(\mathbf{0})), \tag{16a}$$

$$\text{s.t.} \quad \boldsymbol{\lambda} \in \mathcal{D}, \tag{16b}$$

where $U^\ell(\mathbf{0})$ is lower bound for $U^\ell(\boldsymbol{\lambda}^\ell)$, added to ensure the non-negativity of the objective. [4] Indexing flows by both type t and features $\boldsymbol{x}$ implies that, to implement a solution $\boldsymbol{\lambda} \in \mathcal{D}$, routing decisions at intermediate nodes should be based on both quantities. Problem (16) is non-convex in general.[5] Nevertheless, we construct a polynomial time approximation algorithm in the next section.

5 Main Results

Our main contribution is to show that there exists a polynomial-time randomized algorithm that solves Prob. (16) within a $1 - 1/e$ approximation ratio. We do so by establishing that the objective function in Eq. (16a) is *continuous DR- submodular* (see Definition 1).

4. Non-negativity is needed to state guarantees in terms of an approximation ratio (c.f. Thm. 2).

5. It is easy to construct instances of objective (16) that are non-concave. For example, when $|\mathcal{L}| = 1$, $d = 1$, $\mathcal{X} = \{0.1618, 0.3116\}$, $\sigma = 0.0422$, and $\Sigma_\ell = 0.2962$, the Hessian matrix is not negative semi-definite.

Algorithm 1: Frank-Wolfe Variant

Input: $U : \mathcal{D} \to \mathbb{R}_+$, $\mathcal{D}$, stepsize $\delta \in (0, 1]$.
$\boldsymbol{\lambda}^0 = 0, \tau = 0, k = 0$
while $\tau < 1$ **do**
  find $\boldsymbol{v}^k$ s.t. $\boldsymbol{v}^k = \arg\max_{\boldsymbol{v}\in\mathcal{D}} \langle \boldsymbol{v}, \widehat{\nabla U(\boldsymbol{\lambda}^k)} \rangle$
  $\gamma_k = \min\{\delta, 1 - \tau\}$
  $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \gamma_k \boldsymbol{v}^k, \tau = \tau + \gamma_k, k = k + 1$
return $\boldsymbol{\lambda}^K$

5.1 Continuous DR-submodularity

Our first main result establishes the continuous DR-submodularity of the objective (16a):

Theorem 1. *The objective function $U(\boldsymbol{\lambda})$ given by* (16a) *is monotone increasing and continuous DR-submodular in $\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{X}|\times|\mathcal{L}|}$. Moreover,*

$$\frac{\partial U}{\partial \lambda^\ell_{\boldsymbol{x}}} = T \sum_{n=0}^{\infty} \Delta^\ell_{\boldsymbol{x}}(\boldsymbol{\lambda}^\ell, n) \Pr[n^\ell_{\boldsymbol{x}} = n], \tag{17}$$

where $\boldsymbol{n}^\ell$ is distributed as in Eq. (14) *and $\Delta^\ell_{\boldsymbol{x}}(\boldsymbol{\lambda}^\ell, n)$ is:*

$$\mathbb{E}\left[G^\ell(\boldsymbol{n}^\ell) | n^\ell_{\boldsymbol{x}} = n + 1\right] - \mathbb{E}\left[G^\ell(\boldsymbol{n}^\ell) | n^\ell_{\boldsymbol{x}} = n\right] > 0.$$

The proof can be found in Section 5.3; we establish the positivity of the gradient and non-positivity of the Hessian of U. We note that Theorem 1 identifies a *new type of continuous relaxation to DR-submodular functions*, via Poisson sampling; this is in contrast to the multilinear relaxation [27], [29], [32], which is ubiquitous in the literature and relies on Bernoulli sampling. Finally, though our objective is monotone and continuous DR-submodular, the constraint set $\mathcal{D}$ is *not* down-closed. Hence, the analysis by Bian et al. [15] does not directly apply, while using projected gradient ascent [29] would only yield a $1/2$ approximation guarantee.

5.2 Algorithm and Theoretical Guarantee

Our algorithm is summarized in Algorithm 1. We follow the Frank-Wolfe variant for monotone DR-submodular function maximization by Bian et al. [15], deviating both in the nature of the constraint set $\mathcal{D}$, but also, most importantly, in the way we estimate the gradients of objective U.

Frank-Wolfe Variant. In the proposed Frank-Wolfe variant, variables $\boldsymbol{\lambda}^k$ and $\boldsymbol{v}^k$ denote the solution and update direction at the k-th iteration, respectively. Starting from $\boldsymbol{\lambda}^0 = \mathbf{0} \in \mathcal{D}$, the algorithm iterates as follows:

$$\boldsymbol{v}^k = \arg\max_{\boldsymbol{v}\in\mathcal{D}} \langle \boldsymbol{v}, \widehat{\nabla U(\boldsymbol{\lambda}^k)} \rangle, \tag{18a}$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \gamma_k \boldsymbol{v}^k, \tag{18b}$$

where $\gamma_k \in (0, 1]$ is the stepsize with which we move along direction $\boldsymbol{v}^k$, and $\widehat{\nabla U(\cdot)}$ is an estimator of the gradient ∇U w.r.t. $[\lambda^\ell_{\boldsymbol{x}}]_{\boldsymbol{x}\in\mathcal{X},\ell\in\mathcal{L}}$. The step size is set to $\delta > 0$ for all but the

last step, where it is selected so that the total sum of step sizes equals 1.

We note that we face two challenges preventing us from computing the gradient of ∇U directly via. Eq. (17): (a) the gradient computation involves an infinite summation over $n \in \mathbb{N}$, and (b) conditional expectations in $\Delta^{\ell}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\ell}, n)$ require further computing exponential sums in $|\mathcal{X}|$. Using (17) directly in Algorithm 1 would thus not yield a polynomial-time algorithm. To that end, we replace the gradient $\nabla U(\boldsymbol{\lambda}^k)$ used in the standard Frank-Wolfe method by an estimator, which we describe next.

Gradient Estimator. Our estimator addresses challenge (a) above by truncating the infinite sum, and (b) via sampling. In particular, for $n' \geq \lambda^{\ell}_{\boldsymbol{x}} T$, we estimate partial derivatives via the partial summation:

$$\widehat{\frac{\partial U}{\partial \lambda^{\ell}_{\boldsymbol{x}}}} \equiv T \sum_{n=0}^{n'} \widehat{\Delta^{\ell}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\ell}, n)} \Pr[n^{\ell}_{\boldsymbol{x}} = n]. \tag{19}$$

where estimate $\widehat{\Delta^{\ell}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\ell}, n)}$ is constructed via sampling as follows. At each iteration, we generate N samples $\boldsymbol{n}^{\ell,j}$, $j = 1, \ldots, N$ of the random vector $\boldsymbol{n}^{\ell}$ according to the Poisson distribution in Eq. (14), parameterized by the current solution vector $\boldsymbol{\lambda}^{\ell}$. We then compute the empirical average:

$$\widehat{\Delta^{\ell}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\ell}, n)} = \frac{1}{N} \sum_{j=1}^{N} \left(G^{\ell}\left(\boldsymbol{n}^{\ell,j}|_{n^{\ell,j}_{\boldsymbol{x}} = n+1} \right) - G^{\ell}\left(\boldsymbol{n}^{\ell,j}|_{n^{\ell,j}_{\boldsymbol{x}} = n} \right) \right), \tag{20}$$

where $\boldsymbol{n}^{\ell,j}|_{n^{\ell,j}_{\boldsymbol{x}} = n}$ is equal to vector $\boldsymbol{n}^{\ell,j}$ with $n^{\ell,j}_{\boldsymbol{x}}$ set to n.

Theoretical Guarantee. Extending the analysis of [15], and using Theorem 1, we show that the Frank-Wolfe variant combined with gradients estimated "well enough" yields a solution within a constant approximation factor from the optimal:

Theorem 2. *Let*

$$\lambda_{\mathrm{MAX}} \equiv \max_{\boldsymbol{\lambda} \in \mathcal{D}} \sum_{\ell \in \mathcal{L}} ||\boldsymbol{\lambda}^{\ell}||_1, \text{ and} \tag{21}$$

$$G_{\mathrm{MAX}} \equiv \max_{\ell \in \mathcal{L}, \boldsymbol{x} \in \mathcal{X}} (G^{\ell}(\boldsymbol{e}_{\boldsymbol{x}}) - G^{\ell}(\boldsymbol{0})), \tag{22}$$

where $e_{\boldsymbol{x}}$ is the canonical basis. Then, for any $0 < \epsilon_0, \epsilon_1 < 1$ and $\epsilon_2 > 0$, there exists a $\delta > 0$ such that Algorithm 1 terminates in at most

$$K = O\big((|\mathcal{X}||\mathcal{L}|T^2\lambda^2_{\mathrm{MAX}} + 2\lambda_{\mathrm{MAX}})G_{\mathrm{MAX}}/\epsilon_2\big)$$

iterations, and uses $n' = O(\lambda_{\mathrm{MAX}} T + \ln \frac{1}{\epsilon_1})$ terms and $N = O(T^2 n' K^2 \ln \frac{|\mathcal{X}||\mathcal{L}|K}{\epsilon_0})$ samples in estimator (19)*, so that with probability $1 - \epsilon_0$, the output solution $\boldsymbol{\lambda}^K \in \mathcal{D}$ satisfies:*

$$U(\boldsymbol{\lambda}^K) \geq (1 - e^{\epsilon_1 - 1}) \max_{\boldsymbol{\lambda} \in \mathcal{D}} U(\boldsymbol{\lambda}) - \epsilon_2. \tag{23}$$

The proof can be found in Section 5.4. Theorem 2 implies that, through an appropriate (but polynomial) selection of the total number of iterations K, the number of terms n' and samples N, we can obtain a solution $\boldsymbol{\lambda}$ that is within $1 - e^{-1} \approx 0.63$ from the optimal. The proof crucially relies on (and exploits) the continuous DR-submodularity of objective U, in combination with an analysis of the quality of our gradient estimator, given by Eq. (19).

5.3 Proof of Theorem 1

By the law of total expectation, we have:

$$U^{\ell}(\boldsymbol{\lambda}^{\ell}) = \sum_{n=0}^{\infty} \mathbb{E}\left[G^{\ell}(\boldsymbol{n}^{\ell}) | n^{\ell}_{\boldsymbol{x}} = n \right] \cdot \frac{(\lambda^{\ell}_{\boldsymbol{x}} T)^t e^{-\lambda^{\ell}_{\boldsymbol{x}} T}}{t!}.$$

Notably, $\frac{\partial U}{\partial \lambda^{\ell}_{\boldsymbol{x}}} = \frac{\partial U^{\ell}(\boldsymbol{\lambda}^{\ell})}{\partial \lambda^{\ell}_{\boldsymbol{x}}}$, for which the following is true:

$$\begin{aligned}\frac{\partial U^{\ell}(\boldsymbol{\lambda}^{\ell})}{\partial \lambda^{\ell}_{\boldsymbol{x}}} &= \sum_{n=0}^{\infty} \mathbb{E}\left[G^{\ell}(\boldsymbol{n}^{\ell}) | n^{\ell}_{\boldsymbol{x}} = n \right] \cdot \left(\frac{n}{\lambda^{\ell}_{\boldsymbol{x}}} - T \right) \frac{(\lambda^{\ell}_{\boldsymbol{x}} T)^n e^{-\lambda^{\ell}_{\boldsymbol{x}} T}}{n!} \\ &= \sum_{n=0}^{\infty} \Delta^{\ell}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\ell}, n) \cdot T \cdot \Pr[n^{\ell}_{\boldsymbol{x}} = n] \geq 0,\end{aligned}$$

where the last inequality is true because G is monotone-increasing (Lemma 1).

Next, we compute the second partial derivatives $\frac{\partial^2 U}{\partial \lambda^{\ell}_{\boldsymbol{x}} \partial \lambda^{\ell'}_{\boldsymbol{x}'}}$. It is easy to see that for $\ell \neq \ell'$, we have

$$\frac{\partial^2 U}{\partial \lambda^{\ell}_{\boldsymbol{x}} \partial \lambda^{\ell'}_{\boldsymbol{x}'}} = 0.$$

For $\ell = \ell'$ and $\boldsymbol{x} = \boldsymbol{x}'$, it holds that $\frac{\partial^2 U}{\partial (\lambda^{\ell}_{\boldsymbol{x}})^2} = \frac{\partial^2 U^{\ell}(\boldsymbol{\lambda}^{\ell})}{\partial (\lambda^{\ell}_{\boldsymbol{x}})^2}$, where

$$\begin{aligned}&\frac{\partial^2 U^{\ell}(\boldsymbol{\lambda}^{\ell})}{\partial (\lambda^{\ell}_{\boldsymbol{x}})^2} = \Delta^{\ell}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\ell}, 0) \cdot T^2 e^{-\lambda^{\ell}_{\boldsymbol{x}} T} + \sum_{n=1}^{\infty} \Delta^{\ell}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\ell}, n) \\ &\left(\frac{(\lambda^{\ell}_{\boldsymbol{x}})^{n-1} T^{n+1}}{(n-1)!} - \frac{(\lambda^{\ell}_{\boldsymbol{x}})^{n} T^{n+2}}{n!} \right) e^{-\lambda^{\ell}_{\boldsymbol{x}} T} \\ &= \sum_{n=0}^{\infty} (\Delta^{\ell}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\ell}, n+1) - \Delta^{\ell}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\ell}, n)) \cdot \Pr[n^{\ell}_{\boldsymbol{x}} = n] T^2 \leq 0,\end{aligned} \tag{24}$$

and the last equality follows from the DR-submodularity of G (Lemma 1).

For $\ell = \ell'$ and $\boldsymbol{x} \neq \boldsymbol{x}'$, it holds that $\frac{\partial^2 U}{\partial \lambda^{\ell}_{\boldsymbol{x}} \partial \lambda^{\ell}_{\boldsymbol{x}'}} = \frac{\partial^2 U^{\ell}(\boldsymbol{\lambda}^{\ell})}{\partial \lambda^{\ell}_{\boldsymbol{x}} \partial \lambda^{\ell}_{\boldsymbol{x}'}}$,

$$\begin{aligned}&\frac{\partial^2 U^{\ell}(\boldsymbol{\lambda}^{\ell})}{\partial \lambda^{\ell}_{\boldsymbol{x}} \partial \lambda^{\ell}_{\boldsymbol{x}'}} = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \Big(\Big(\mathbb{E}\left[G^{\ell}(\boldsymbol{n}^{\ell}) | n^{\ell}_{\boldsymbol{x}} = n+1, n^{\ell}_{\boldsymbol{x}'} = k+1 \right] \\ &- \mathbb{E}\left[G^{\ell}(\boldsymbol{n}^{\ell}) | n^{\ell}_{\boldsymbol{x}} = n, n^{\ell}_{\boldsymbol{x}'} = k+1 \right] \Big) - \Big(\mathbb{E}[G^{\ell}(\boldsymbol{n}^{\ell}) | \\ &n^{\ell}_{\boldsymbol{x}} = n+1, n^{\ell}_{\boldsymbol{x}'} = k] - \mathbb{E}\left[G^{\ell}(\boldsymbol{n}^{\ell}) | n^{\ell}_{\boldsymbol{x}} = n, n^{\ell}_{\boldsymbol{x}'} = k \right] \Big) \Big) \\ &\cdot \Pr[n^{\ell}_{\boldsymbol{x}} = n] \Pr[n^{\ell}_{\boldsymbol{x}'} = k] T^2 \leq 0,\end{aligned} \tag{25}$$

where the last inequality follows from the DR-submodularity of G (Lemma 1). □

5.4 Proof of Theorem 2

Our proof relies on a series of key lemmas; we state them below. Full proofs of all lemmas can be found in the appendix. We begin by associating the approximation guarantee of Algorithm 1 to the quality of gradient estimation $\widehat{\nabla U(\cdot)}$:

Lemma 2. *Suppose we can construct an estimator $\widehat{\nabla U(\boldsymbol{\lambda}^k)}$ of the gradient $\nabla U(\boldsymbol{\lambda}^k)$ at each iteration k such that*

$$\langle \boldsymbol{v}^k, \nabla U(\boldsymbol{\lambda}^k) \rangle \geq a \cdot \max_{\boldsymbol{v} \in \mathcal{D}} \langle \boldsymbol{v}, \nabla U(\boldsymbol{\lambda}^k) \rangle - b, \tag{26}$$

where $\boldsymbol{v}^k$ is the update direction determined by (18a), *$a \in (0,1]$ and b are positive constants. Then, the output solution $\boldsymbol{\lambda}^K$ of Algorithm 1 satisfies $\boldsymbol{\lambda}^K \in \mathcal{D}$, and*

$$U(\boldsymbol{\lambda}^K) \geq (1-e^{-a}) \max_{\boldsymbol{\lambda}\in\mathcal{D}} U(\boldsymbol{\lambda}) - \frac{L}{2}\lambda_{\rm MAX}^2\delta - b, \qquad (27)$$

where $L = 2|\mathcal{X}||\mathcal{L}|T^2G_{\rm MAX}$ is the Lipschitz constant of ∇U, and $\lambda_{\rm MAX}$ and $G_{\rm MAX}$ given by (21) *and* (22).

The proof, found in Appendix B, relies on the continuous DR-submodularity of U, and follows [15]; we deviate from their proof to handle the additional issue that $\mathcal{D}$ is not downward closed (an assumption in [15]).

Next, we turn our attention to characterizing the quality of our gradient estimator. To that end, use the following subexponential tail bound:

Lemma 3 (Theorem 1 in [45])**.** *Let $n_{\boldsymbol{x}}^\ell \sim \text{Poisson}(\lambda_{\boldsymbol{x}}^\ell T)$, for $\lambda_{\boldsymbol{x}}^\ell, T > 0$. Then, for any $z > \lambda_{\boldsymbol{x}}^\ell T$, we have*

$$\Pr[n_{\boldsymbol{x}}^\ell \geq z] \leq e^{-\frac{(z-\lambda_{\boldsymbol{x}}^\ell T)^2}{2\lambda_{\boldsymbol{x}}^\ell T} h(\frac{z-\lambda_{\boldsymbol{x}}^\ell T}{\lambda_{\boldsymbol{x}}^\ell T})}, \qquad (28)$$

where $h : [-1,\infty) \to \mathbb{R}$ is the function defined by $h(u) = 2\frac{(1+u)\ln(1+u)-u}{u^2}$.

The expression for $h(u)$ implies that the Poisson tail decays slightly faster than a standard exponential random variable (by a logarithmic factor). This lemma allows us to characterize the effect of truncating Eq. (17) in estimation quality. In particular, for $n' \geq \lambda_{\boldsymbol{x}}^\ell T$, let:

$$\text{HEAD}_{\boldsymbol{x}}^\ell(n') \equiv T\sum_{n=0}^{n'} \Delta_{\boldsymbol{x}}^\ell(\boldsymbol{\lambda}^\ell, t)\Pr[n_{\boldsymbol{x}}^\ell = n]. \qquad (29)$$

Then, this is guaranteed to be within a constant factor from the true partial derivative:

Lemma 4. *For $h(u) = 2\frac{(1+u)\ln(1+u)-u}{u^2}$ and $n' \geq \lambda_{\boldsymbol{x}}^\ell T$, we have:*

$$\begin{aligned}\frac{\partial U}{\partial \lambda_{\boldsymbol{x}}^\ell} &\geq \text{HEAD}_{\boldsymbol{x}}^\ell(n') \\ &\geq (1-e^{-\frac{(n'-\lambda_{\boldsymbol{x}}^\ell T+1)^2}{2\lambda_{\boldsymbol{x}}^\ell T} h(\frac{n'-\lambda_{\boldsymbol{x}}^\ell T+1}{\lambda_{\boldsymbol{x}}^\ell T})})\frac{\partial U}{\partial \lambda_{\boldsymbol{x}}^\ell}.\end{aligned} \qquad (30)$$

The proof can be found in Appendix C.

Next, by estimating $\Delta_{\boldsymbol{x}}^\ell(\boldsymbol{\lambda}^\ell, n)$ via sampling (see (20)), we construct our final estimator given by (19). Putting together Lemma 4 and along with a Chernoff bound [46], to attain a guarantee on sampling, we can bound the quality of our gradient estimator:

Lemma 5. *At each iteration k, with probability greater than $1 - 2|\mathcal{X}||\mathcal{L}| \cdot e^{-\delta^2 N/2T^2(n'+1)}$,*

$$\langle \boldsymbol{v}^k, \nabla U(\boldsymbol{\lambda}^k)\rangle \geq a \cdot \max_{\boldsymbol{v}\in\mathcal{D}} \langle \boldsymbol{v}, \nabla U(\boldsymbol{\lambda}^k)\rangle - b, \qquad (31)$$

where

$$a = 1 - \max_{k=1,\ldots,K} \text{P}_{\rm MAX}^k, \quad \textit{and} \qquad (32)$$

$$b = 2\lambda_{\rm MAX}\delta \cdot G_{\rm MAX}, \qquad (33)$$

for $\text{P}_{\rm MAX}^k = \max_{l\in\mathcal{L},\boldsymbol{x}\in\mathcal{X}} \Pr[n_{\boldsymbol{x}}^{\ell,k} \geq n'+1]$ ($n_{\boldsymbol{x}}^{\ell,k}$ is a Poisson r.v. with parameter $\lambda_{\boldsymbol{x}}^{\ell,k}T$), and with $\lambda_{\rm MAX}$ and $G_{\rm MAX}$ given by Eq. (21) *and* (22).

The proof is in Appendix D.

Theorem 2 follows by combining Lemmas 2 and 5. In particular, by Lemma 5 and a union bound, we have that (31) is satisfied for all iterations with probability greater than $1 - 2|\mathcal{X}||\mathcal{L}|K \cdot e^{-\delta^2 N/2T^2(n'+1)}$. This, combined with Lemma 2, implies that

$$\begin{aligned}U(\boldsymbol{\lambda}^K) \geq &(1-e^{\text{P}_{\rm MAX}-1}) \cdot \max_{\boldsymbol{\lambda}\in\mathcal{D}} U(\boldsymbol{\lambda}) \\ &- (|\mathcal{X}||\mathcal{L}|T^2\lambda_{\rm MAX}^2 + 2\lambda_{\rm MAX})G_{\rm MAX}\delta,\end{aligned}$$

is satisfied with the same probability. This implies that for any $0 < \epsilon_0, \epsilon_1 < 1$ and $\epsilon_2 > 0$,

$$U(\boldsymbol{\lambda}^K) \geq (1-e^{\epsilon_1-1}) \cdot \text{OPT} - \epsilon_2,$$

with probability $1-\epsilon_0$. From Eq. (28), the probability is an increasing function w.r.t. $\lambda_{\boldsymbol{x}}^\ell$, and $\lambda_{\rm MAX}$ is an upper bound for $\lambda_{\boldsymbol{x}}^\ell$. Letting

$$u = \frac{n' - \lambda_{\rm MAX}T}{\lambda_{\rm MAX}T},$$

we have

$$\begin{aligned}\text{P}_{\rm MAX} &< e^{-\frac{(n'-\lambda_{\rm MAX}T)^2}{2\lambda_{\rm MAX}T} h(\frac{n'-\lambda_{\rm MAX}T}{\lambda_{\rm MAX}T})} \\ &= e^{-\lambda_{\rm MAX}T((1+u)\ln(1+u)-u)} \leq \Omega(e^{-\lambda_{\rm MAX}Tu}) = \epsilon_1,\end{aligned}$$

where the last inequality holds because $(1+u)\ln(1+u) - u > u$ when u is large enough, e.g., $u \geq 4$. Thus, $n' = O(\lambda_{\rm MAX}T + \ln\frac{1}{\epsilon_1})$.
We determine K and N by setting

$$(|\mathcal{X}||\mathcal{L}|T^2\lambda_{\rm MAX}^2 + 2\lambda_{\rm MAX})G_{\rm MAX}/K = \epsilon_2$$

and

$$2|\mathcal{X}||\mathcal{L}|K \cdot e^{-N/2T^2(n'+1)K^2} = \epsilon_0.$$

Therefore, $K = O((|\mathcal{X}||\mathcal{L}|T^2\lambda_{\rm MAX}^2 + 2\lambda_{\rm MAX})G_{\rm MAX}/\epsilon_2)$, and $N = O(T^2n'K^2\ln\frac{|\mathcal{X}||\mathcal{L}|K}{\epsilon_0})$. □

6 EXTENSIONS

Our model extends in many ways (e.g., to multiple types per learner). We discuss three non-trivial extensions below.

Heterogeneous Noisy Sources. Our model and analysis directly generalizes to a heterogeneous (or *heteroskedastic*) noise setting, in which the noise level varies across sources. Formally, labels of type t at source s are generated via $y = \boldsymbol{x}^\top\boldsymbol{\beta}_t + \epsilon_{t,s}$, where $\epsilon_{t,s}$ are zero-mean normal noise variables with variance $\sigma_{t,s}^2$. In this case, the estimator in (2) needs to be replaced by Generalized Least Squares [47], whereby every pair $(\frac{\boldsymbol{x}}{\sigma_t}, \frac{y}{\sigma_t}) \in \mathbb{R}^d \times \mathbb{R}$ of type t generated by s is replaced by $(\frac{\boldsymbol{x}}{\sigma_{t,s}}, \frac{y}{\sigma_{t,s}}) \in \mathbb{R}^d \times \mathbb{R}$ prior to applying Eq. (2). This, in turn, changes the D-optimality criterion objective, so that σ_t^2 is replaced by $\sigma_{t,s}^2$ for vectors $\boldsymbol{x} \in \mathcal{X}$ coming from source s. In other words, data coming from noisy sources are valued less by the learner. This rescaling preserves the monotonicity and continuous DR-submodularity of our overall objective, and our guarantees hold, *mutatis mutandis*.

Uncountable $\mathcal{X}$. We assumed in our analysis that data features are generated from a finite set $\mathcal{X}$, and that transmission rates per edge are parametrized by both the type $t \in \mathcal{T}$ *and* the features $\boldsymbol{x}$ of the data pair transmitted. This a priori prevents us from considering an infinite set

of experiments $\mathcal{X}$: this would lead to an infinite set of constraints in Problem (16). In practice, it would also make routing intractable, as routing decisions depend on both t and $\boldsymbol{x}$.

We can however extend our analysis to a setting where experiments $\mathcal{X}$ are infinite, or even uncountable. To do so, we can consider rates per edge e of the form $\lambda^e_{s,t}$, i.e., parameterized by type t and *source* s rather than features $\boldsymbol{x}$. In practice, this would mean that packets would be routed based on the source and type, not inspecting the features of the internal pairs, while constraints would be finite (depending on $|\mathcal{S}|$, not $|\mathcal{X}|$). Data generation at source s can then be modelled via a compound Poisson process with rate $\lambda_{s,t}$, at the epochs of which the features $\boldsymbol{x}$ are sampled independently from some probability distribution $\nu_{s,t}$ over $\mathbb{R}^d$. The objective then would be written as an expectation over not only arrivals at a learner from source s (which will again be Poisson) but also the distribution ν_{s,t^ℓ} of features. Sampling from the latter would need to be used when estimating ∇U; as long as Chernoff-type bounds can be used to characterize the estimation quality of such sampling (which would be the case if, e.g., $\nu_{s,t}$ are Gaussian), our analysis would still hold, taking also the number of sampled features into account.

No-Label Setting. The labeled setting is interesting, as sensors can very well be collecting covariates as well as a target variable (e.g., temperature) that is to be regressed from these covariates. Nevertheless, we would like to point out that our entire setting (model, objective, and algorithms) *can also be used in the label-less setting directly, without any modifications.* In particular, observe that the optimization problem we are solving (Eq. (16)), only uses the feature vectors $\boldsymbol{x}$, both in terms of the objective, how variables are described, as well as the constraints. Labels y are only pertinent in mapping sources to learners (i.e., a learner interested in a type would only get data pertaining to that type). The actual labels themselves are *not used anywhere other than during training*: they appear neither in the optimization nor in the resulting solution (i.e., the routing and scheduling scheme.) In practice, this means that the algorithms we proposed can be used in a label-less setting, where sensors do not collect labels. In this setting, learners obtain unlabeled data (fully described by features $\boldsymbol{x}$) and then label them in place, either manually or otherwise. I.e., the training still happens at labelled data, where the labeling happens at a later time (after the data has been delivered). Our optimization would be exactly the same, determining how samples should be routed.

Arbitrary (Non-DAG) Topology. For notational convenience, we assumed that graph G was a DAG, with sources and sinks corresponding to sets $\mathcal{S}$ and $\mathcal{L}$ respectively. Our analysis further extends to more general (i.e., non-DAG) graphs, provided that extra care is taken for flow constraints to prevent cycles. This can be accomplished, e.g., via source routing. Given an arbitrary graph, and arbitrary locations for data sources and learners, we can extend our setting as follows: (a) flows from a source s to a learner ℓ could follow source-path routing, over one or more directed paths linking the two, and (b) flows could be indexed by (and remain constant along) a path, in addition to $\boldsymbol{x}$ and t, while also ensuring that (c) aggregate flow across all paths that pass through an edge does not violate capacity constraints. Such a formulation still yields a linear set of constraints, and our analysis still holds. In fact, in this case, the corresponding set $\mathcal{D}$ is downward closed, so the proof of the corresponding Theorem 2 follows more directly from [15].

7 Numerical Evaluation

In this section, we provide a comprehensive evaluation of the Frank-Wolfe variant algorithm to understand how rate allocation affects model learning quality.

7.1 Experimental Setting

To evaluate the proposed algorithm, we experiment [6] with three distinct settings (Setting1, Setting2, and Setting3). In all three settings, we consider a finite feature vector set $\mathcal{X}$ that includes randomly generated feature vectors with $d = 100$, and a set $\mathcal{T}$ that of different Bayesian linear regression models with $\boldsymbol{\beta}_t$, $t \in \mathcal{T}$. Labels of each type are generated with Gaussian noise, whose variance σ_t is uniformly at random (u.a.r.) chosen from 0.5 to 1. For each network, we u.a.r. select $|\mathcal{L}|$ learners and $|\mathcal{S}|$ data sources, and remove incoming edges of sources and outgoing edges of learners. Each learner has a target model $\boldsymbol{\beta}_{t^\ell}$, $t^\ell \in \mathcal{T}$ with a diagonal prior $\boldsymbol{\Sigma}_{t^\ell}$ generated as follows. First, we separate features into two classes: well-known and poorly-known. Then, we set the corresponding prior covariance (i.e., the diagonal elements in $\boldsymbol{\Sigma}_{t^\ell}$) to low (uniformly from 0 to 0.01) and high (uniformly from 100 to 200) values, for well-known and poorly-known features, respectively. The link capacity $\mu^e, e = (u, v) \in \mathcal{E}$ is selected u.a.r. from 20 to 50.

In Setting1, we set $|\mathcal{S}|$ and $|\mathcal{L}|$ as shown in Table 2. Moreover, each source s generates data $(\boldsymbol{x}, y)$ of type t label with rate $\lambda^s_{\boldsymbol{x},t}$, uniformly distributed over [2,5]. In Setting2, all sources are homogeneous, i.e., $\lambda^s_{\boldsymbol{x},t} = 5$. Finally, in Setting3, we conduct experiments with a much larger set of sources and learners ($|\mathcal{S}|$ and $|\mathcal{L}|$, respectively), as indicated in Table 2.

7.2 Algorithms

We compare our proposed Frank-Wolfe based algorithm (we denote it by `FW`) with several baseline data transmission strategies derived in different ways:

- `MaxSum`: This maximizes the aggregate total useful incoming traffic rates of learners, i.e., the objective is:

$$U_{\texttt{MaxSum}}(\boldsymbol{\lambda}) = \sum_{\ell \in \mathcal{L}} \sum_{\boldsymbol{x} \in \mathcal{X}} \lambda^\ell_{\boldsymbol{x}}.$$

- `MaxAlpha`: This maximizes the aggregate α-fair utilities [50] of the total useful incoming traffic at learners, i.e., the objective is:

$$U_{\texttt{MaxAlpha}}(\boldsymbol{\lambda}) = \sum_{\ell \in \mathcal{L}} (\sum_{\boldsymbol{x} \in \mathcal{X}} \lambda^\ell_{\boldsymbol{x}})^{1-\alpha}/(1-\alpha).$$

  We set $\alpha = 5$.

We also compare with another algorithm for the proposed experimental design networks:

6. Our code and data are publicly available at https://github.com/neu-spiral/Networked-Learning.

TABLE 2
Graph Topologies and Experiment Parameters

Graph	$\|V\|$	$\|E\|$	$\|\mathcal{X}\|$	Setting1/2 $\|\mathcal{T}\|$	Setting1/2 $\|\mathcal{S}\|$	Setting1/2 $\|\mathcal{L}\|$	Setting3 $\|\mathcal{T}\|$	Setting3 $\|\mathcal{S}\|$	Setting3 $\|\mathcal{L}\|$
synthetic topologies									
Erdős-Rényi (ER)	100	1042	20	5	10	5	10	20	15
balanced-tree (BT)	341	680	20	5	10	5	10	20	15
hypercube (HC)	128	896	20	5	10	5	10	20	15
star	100	198	20	5	10	5	10	20	15
grid	100	360	20	5	10	5	10	20	15
small-world (SW) [48]	100	491	20	5	10	5	10	20	15
real backbone networks [49]									
GEANT	22	66	20	3	3	3	4	4	4
Abilene	9	26	20	3	3	3	4	4	4
Deutsche Telekom (Dtelekom)	68	546	20	3	3	3	4	4	4

- PGA: it also solves Prob. (16), as does our proposed algorithm, via the projected gradient ascent [29]. As PGA also requires gradients, we use our novel gradient estimation (by Eq. (19)).

Note that projected gradient ascent finds a solution within $1/2$ from the optimal if the true gradients are accessible [29] (a theoretical guarantee with estimated gradients is out of our scope). We run FW and PGA for $K = 50$ iterations with step size $\delta = 0.02$. In each iteration, we estimate the gradient according to Eq. (19) with $N = 50$, and $n' = \lceil 2 \max_{\ell,\boldsymbol{x}} \lambda_{\boldsymbol{x}}^{\ell} T \rceil$, where $\lambda_{\boldsymbol{x}}^{\ell}$ is given by the current solution. We consider a data acquisition time $T = 1$.

7.3 Performance Metrics

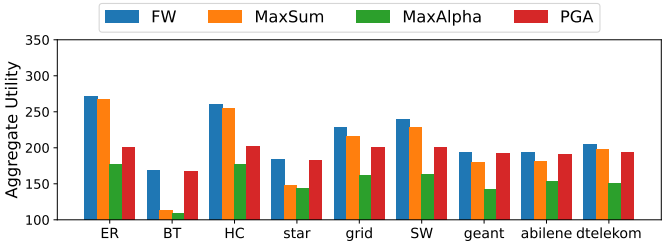
To evaluate the performance of algorithms, we use *Aggregate Utility*, defined in (16a) as one metric. Note that, aggregate utility involves summation with infinite support, we thus need to resort to sampling to estimate it, where we sample with 1000 samples. Also, we define an *Average Norm of Estimation Error* to measure the model estimation quality. Formally, it is defined as:

$$\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\|\hat{\boldsymbol{\beta}}_{\texttt{MAP}}^{\ell} - \boldsymbol{\beta}^{\ell}\| = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\|(\boldsymbol{X}^{\ell\top}\boldsymbol{X}^{\ell} + \sigma_{i^{\ell}}^{2}\boldsymbol{\Sigma}_{\ell}^{-1})^{-1}\cdot(\boldsymbol{X}^{\ell\top}\boldsymbol{y}^{\ell} + \sigma_{i^{\ell}}^{2}\boldsymbol{\Sigma}_{\ell}^{-1}\boldsymbol{\beta}_{0}^{\ell}) - \boldsymbol{\beta}^{\ell}\|, \tag{34}$$

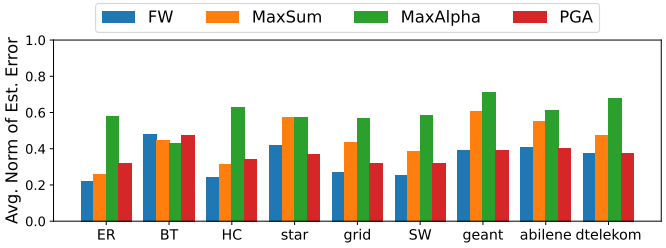
following the equation of MAP estimation (see Eq. (2)). We average over 1000 realizations of the label noise as well as the number of data arrived at the learner $\{\boldsymbol{n}^{\ell}\}_{\ell\in\mathcal{L}}$ to calculate this expectation.

7.4 Results

Performance over Different Topologies. We first compare the proposed algorithm (FW) with baselines in terms of the aggregated utility and model estimation quality over several networks, shown in Figures 2(a) and 2(b), respectively (Setting1). Learners in these networks have distinct target models to train. In all network topologies, FW outperforms MaxSum and MaxAlpha in both aggregate utility and average norm of estimation error. PGA, which also is based on our experimental design framework, performs well (second best) in most networks w.r.t. estimation quality. In experiments of synthetic topologies, two learners may learn



(a) Aggregate Utility



(b) Average Norm of Estimation Error

Fig. 2. Aggregate utility and average norm of estimation error in different networks. We can observe that FW is the best in terms of maximizing the utility and minimizing the estimation error in all networks.

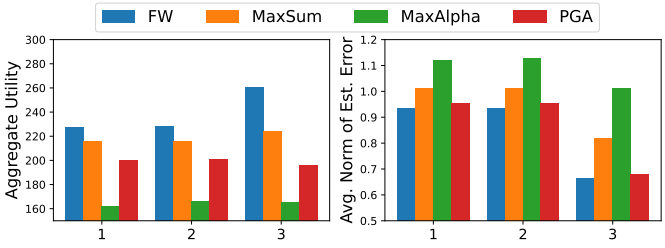
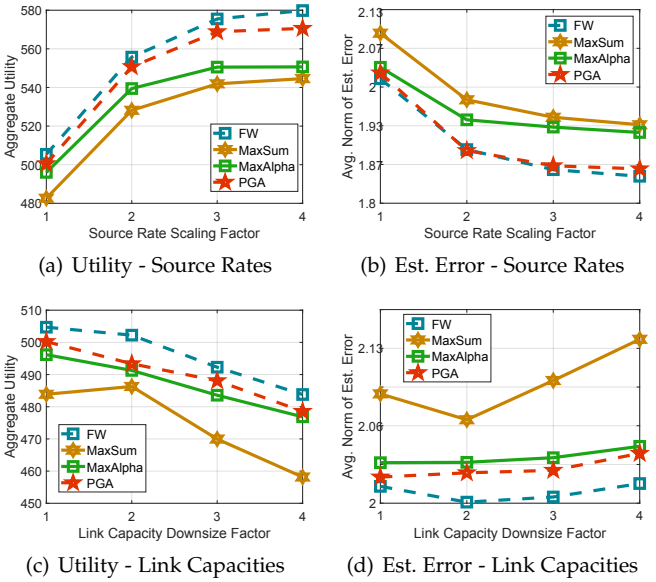


Fig. 3. Aggregate utility and average norm of estimation error per learner of grid topology for three settings. All settings have similar result pattern: FW and PGA perform better than other two, and FW is slightly better than PGA.

the same type of model. The results patterns of Setting2 and Setting3 are similar to Setting1, where FW outperforms other competitors. This also verifies Thm. 2 experimentally: arrival rates, the number of learners, the number of sources

TABLE 3
Aggregate Utility for Different Settings. Our algorithm FW outperforms the competitors in all settings.

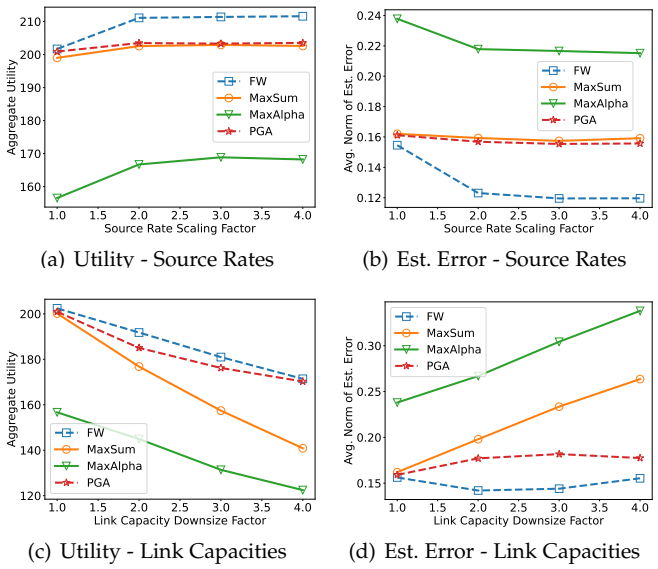
Graph	Setting1				Setting2				Setting3			
	$U_{\texttt{FW}}$	$U_{\texttt{MaxSum}}$	$U_{\texttt{MaxAlpha}}$	$U_{\texttt{PGA}}$	$U_{\texttt{FW}}$	$U_{\texttt{MaxSum}}$	$U_{\texttt{MaxAlpha}}$	$U_{\texttt{PGA}}$	$U_{\texttt{FW}}$	$U_{\texttt{MaxSum}}$	$U_{\texttt{MaxAlpha}}$	$U_{\texttt{PGA}}$
Erdős-Rényi (ER)	270.8	266.9	177.3	200.5	270.7	266.6	188.3	200.7	391.6	378.5	218.3	195.3
balanced-tree (BT)	167.9	112.8	109.4	167.3	168.9	110.7	111.6	168.7	207.3	141.2	123.0	187.0
hypercube (HC)	260.7	253.9	176.9	201.3	260.7	254.1	174.2	201.7	368.3	354.6	199.9	191.0
star	184.4	147.9	144.1	182.7	184.7	147.8	141.4	182.5	217.3	155.9	149.2	196.1
grid	227.5	215.6	161.9	200.0	227.7	215.3	165.9	200.4	260.0	224.1	164.9	196.0
small-world (SW) [48]	238.9	228.2	163.6	200.6	238.7	228.5	172.5	200.8	293.8	269.5	173.3	196.3
GEANT	193.3	179.3	141.4	191.6	193.4	178.4	140.0	192.2	170.3	125.5	125.6	168.9
Abilene	193.8	181.0	152.6	190.0	195.1	180.2	149.5	194.3	189.5	174.9	125.3	181.3
Deutsche Telekom (Dtelekom)	205.1	197.8	150.1	193.8	208.2	196.8	153.0	193.9	212.9	202.1	158.6	191.1



(a) Utility - Source Rates (b) Est. Error - Source Rates

(c) Utility - Link Capacities (d) Est. Error - Link Capacities

Fig. 4. Algorithm comparison in abilene topology where two of the learners have a same target model and the third has a different one. The achieved aggregate utility and model estimation quality of different algorithms are evaluated with different source data rates and bottleneck link capacities.



(a) Utility - Source Rates (b) Est. Error - Source Rates

(c) Utility - Link Capacities (d) Est. Error - Link Capacities

Fig. 5. Algorithm comparison in Abilene topology with different source rates and link capacities where all learners have different target models.

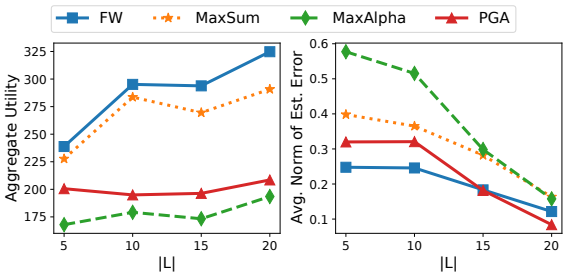
and to name a few, all these do not affect our optimality guarantees. We take grid topology as an example and show the results in Fig. 3. We also list aggregate utility of all topologies in three settings in Tab. 3 for reference.

Effect of Source Rates and Link Capacities. Next, we evaluate how algorithm performance is affected by varying source rates as well as link capacities. We focus on the Abilene network, having 3 sources and 3 learners, the same as that in Setting1/2, instead two of the learners have a same target training model.

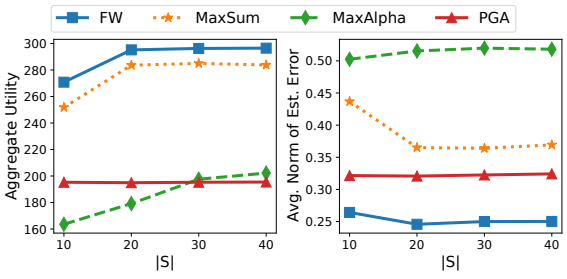
Figures 4(a) and 4(b) plot the aggregate utility and average total norm of estimation error across learners, with different data source rates at the sources. The initial source rates are sampled u.a.r. from 2 to 5, and we scale it by different scaling factors. As the source rates increases, the aggregate utility increases and the average norm of estimation error decreases for all algorithms. FW is always the best in both figures. Moreover, the proposed experimental design framework can significantly improve the training quality: algorithms based on our proposed framework (FW and PGA) with source rates scaled by 2 already outperform the other two algorithms (MaxSum and MaxAlpha) with source rates scaled by 4. We see reverse results of MaxSum and MaxAlpha in these two figures compared with Figure 2, showing that the algorithm which considers fairness (i.e., α-fair utilities), may perform better if we have competing learners.

Figures 4(c) and 4(d) show performance in Abilene network with different link capacities of several bottleneck links. The initial link capacities are divided by different downsize factors. The overall trend is that as the link capacities decrease, algorithms achieve smaller aggregate network utility and get a higher average norm of estimation error. The proposed algorithm is always the best with different bottleneck link capacities in both figures.

In Fig. 5, we further consider the same Abilene network as in Fig. 2 and the same settings where all learners have different target models, varying source rates and link capacities. Compared with Fig. 4, we see similar trends for all algorithms. The difference is that the MaxAlpha performs much worse than MaxSum and other algorithms gaining fewer benefits from considering fairness. The purposed FW outperforms other baselines with all different parameter set-



(a) Effect of Learners Size



(b) Effect of Sources Size

Fig. 6. Algorithms comparison in SW topology with different learners size and sources size. Aggregate utility increases, while average norm of estimation error decreases, with more learners and sources.

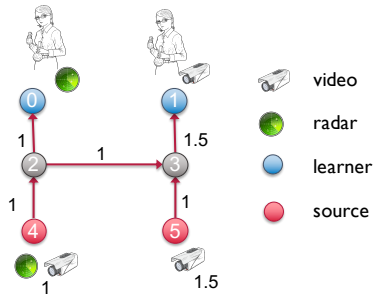


Fig. 7. A simple illustrative example for network design. There are two learners (nodes 0 and 1) and two sources (nodes 4 and 5). Link capacities are shown next to edges, and source generation rates are shown next to sources. The two types are also indicated in the figure. As rates in source 1 exceed the capacities of edge (4,2), the latter is a bottleneck edge, and a design needs to make intelligent use of this resource.

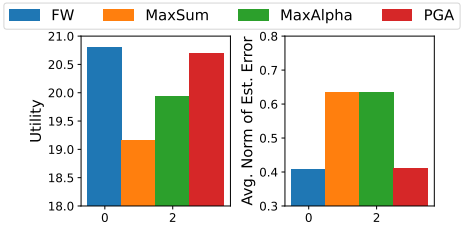


Fig. 8. Aggregate utility and average norm of estimation error in a simple example. Our proposed FW outperform all the competitors.

tings, showing the advantage of our purposed framework.

Effect of Source and Learner Set Size. Finally, the impact of source set size $|\mathcal{S}|$ and learner set size $|\mathcal{L}|$ on aggregate utility and average norm of estimation error is shown in Fig. 6, where we take SW as an example. A comparison in Tab. 3 and Fig. 3 between Setting1 and Setting3 also reflects similar phenomena. To explore the effect of the learner set size, we keep the source set size $|\mathcal{S}| = 20$ and types set size $|\mathcal{T}| = 10$ for Fig. 6(a). The figure shows that when increasing the number of learners, the aggregate utility increases and estimation error decreases. This is because D-optimal design is a DR-submodular function, which has a similar effect as concavity: if the total data arrival is fixed, even distribution of the data features leads to larger aggregate utility.

We study the impact of source set size by setting $|\mathcal{L}| = 10$ and $|\mathcal{T}| = 10$ for Fig. 6(b). As the number of sources increases, the aggregate utility first increases very fast and then tapers off. More sources reflect greater data generation rates, and thus greater utility. However, because of DR-submodularity, the marginal gain decreases as the number of sources increases. Furthermore, under limited bandwidth, if useful feature vectors reach saturation, there will be little utility increment as well. The same interpretation applies to the estimation error in the opposite direction.

7.5 An Illustrative Simple Example

To further illustrate how the Frank-Wolfe Variant and our particular objective behave we explore its performance on a simple network design problem, shown in Fig. 7. There are two types of labels: "video", denoted by type 0, and "radar", denoted by type 1. Node 0 is a learner targeting type "radar", and node 1 is a learner targeting type "video". Each has a distinct prior covariance, with $d = 10$, generated as described in Sec. 7.1, split in two: low confidence features for learner 0 are high confidence features for learner 1, and vice-versa.

Node 4 is a source which generates radar and video data with rates equal to 1, respectively, and node 5 is a source which generates video data only with rate 1.5. Link capacities of each edge are shown in Fig. 7. We consider a set $\mathcal{X}$ with two feature vectors with dimension $d = 10$: $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$. These two vectors have distinct supports, split exactly the same way as learners: hence, $\boldsymbol{x}_0$ is valuable to learner 0, and $\boldsymbol{x}_1$ is valuable to learner 1.

Edge (4,2) has link capacity 1, which is half than the rate with which source 2 generates data. As a result, this is a bottleneck edge, and the network needs to intelligently allocate allocate rates across the two different types, so that the two learners indeed learn their respective model better.

Fig. 9 displays the rate allocated by different algorithms for this instance. Algorithms FW and PGA have the same objectives, so that the rates allocated look very similar. In particular, in both allocations, learner 0 prefers features 0 with type 1 and learner 1 prefers features 1 with type 0. Algorithms MaxSum and MaxAlpha only focus on the aggregate traffic, and do not distinguish different features and types. This is verified the span of the allocation support in MaxSum and MaxAlpha, e.g., over edge (3,1) and (5,3), and also the same rates over the edge with different feature vectors. MaxAlpha considers fairness over learner, which is reflected by equal arrival rates at each learner.

We also compare the aggregate utility and model estimation quality for all algorithms, shown in Fig. 8. Algorithms FW and PGA have similar performance, much better than MaxAlpha and MaxAlpha. Our proposed FW performs the

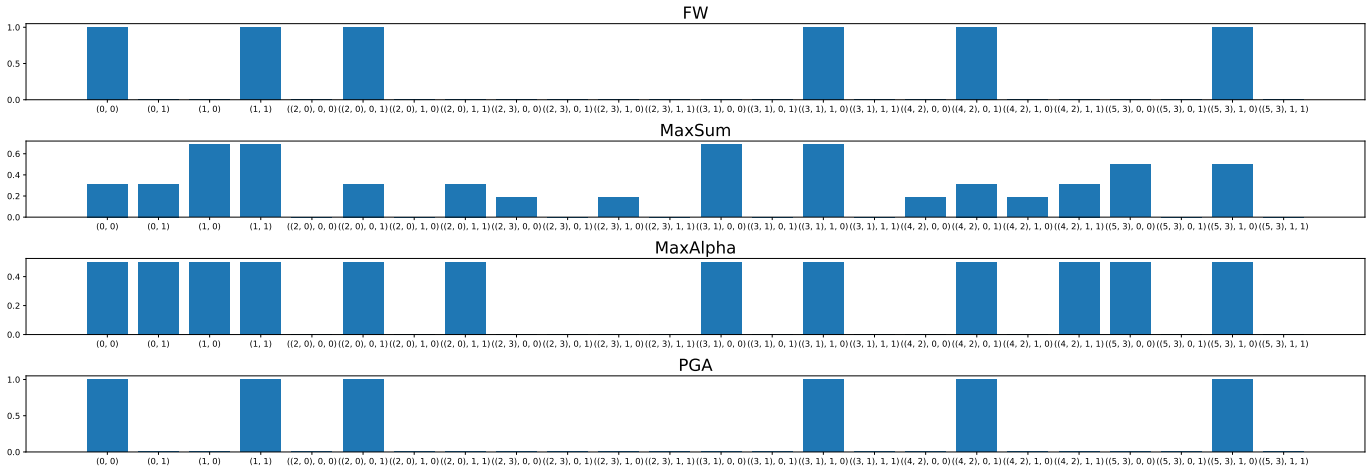


Fig. 9. Rate allocation over learners and edges for different algorithms. The first four columns indicate the rates at (learner, feature vector). The rest indicates (edge, feature vector, type). Algorithms `FW` and `PGA` have preferred features with targeted type, while `MaxSum` and `MaxAlpha` do not.

best along all the algorithms.

8 Conclusion

We propose *experimental design networks*, to determine a data transmission strategy that maximizes the quality of trained models in a distributed learning system. The underlying optimization problem can be approximated even though its objective function is non-concave.

Beyond extensions we have already listed, our framework can be used to explore other experimental design objectives (e.g., A-optimality and E-optimality) as well as variants that include data source generation costs. Distributed and adaptive implementations of the rate allocation schemes we proposed are also interesting future directions. Incorporating non-linear learning tasks (e.g., deep neural networks) is also an open avenue of exploration: though Bayesian posteriors are harder to compute in closed-form for this case, techniques such as variational inference [51] could be used in this case. In turn, these could be used to model the contribution of data to the reduction in model uncertainty. Finally, an interesting extension of our model involves a multi-stage setting, in which learners receive data in one stage, update their posteriors, and use these as new priors in the next stage. Studying the dynamics of such a system, as well as how network design impacts these dynamics, is a very interesting open problem.

Acknowledgment

The authors gratefully acknowledge support from the National Science Foundation (grants 1718355, 2107062, and 2112471).

References

[1] Y. Liu, Y. Li, L. Su, E. Yeh, and S. Ioannidis, "Experimental design networks: A paradigm for serving heterogeneous learners under networking constraints," *arXiv preprint*, 2022.

[2] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2017.

[3] B. Yang, X. Cao, X. Li, Q. Zhang, and L. Qian, "Mobile-edge-computing-based hierarchical machine learning tasks distribution for iiot," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 2169–2180, 2019.

[4] V. Albino, U. Berardi, and R. M. Dangelico, "Smart cities: Definitions, dimensions, performance, and initiatives," *Journal of urban technology*, vol. 22, no. 1, pp. 3–21, 2015.

[5] M. Mohammadi and A. Al-Fuqaha, "Enabling cognitive smart cities using big data and machine learning: Approaches and challenges," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 94–101, 2018.

[6] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.

[7] Y. Deshpande and A. Montanari, "Linear bandits in high dimension and recommendation systems," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 1750–1754.

[8] B. Settles, *Active Learning Literature Survey*. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[9] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data lifecycle challenges in production machine learning: a survey," *ACM SIGMOD Record*, vol. 47, no. 2, pp. 17–28, 2018.

[10] T. Horel, S. Ioannidis, and S. Muthukrishnan, "Budget feasible mechanisms for experimental design," in *Latin American Symposium on Theoretical Informatics*. Springer, 2014, pp. 719–730.

[11] Y. Guo, J. Dy, D. Erdogmus, J. Kalpathy-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, and S. Ioannidis, "Accelerated experimental design for pairwise comparisons," in *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 2019, pp. 432–440.

[12] N. Gast, S. Ioannidis, P. Loiseau, and B. Roussillon, "Linear regression from strategic data sources," *ACM Transactions on Economics and Computation (TEAC)*, vol. 8, no. 2, pp. 1–24, 2020.

[13] Y. Guo, P. Tian, J. Kalpathy-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, D. Erdogmus, J. G. Dy, and S. Ioannidis, "Experimental design under the bradley-terry model." in *IJCAI*, 2018, pp. 2198–2204.

[14] P. Flaherty, A. Arkin, and M. I. Jordan, "Robust design of biological experiments," in *Advances in Neural Information Processing Systems*, 2006, pp. 363–370.

[15] A. A. Bian, B. Mirzasoleiman, J. Buhmann, and A. Krause, "Guaranteed non-convex optimization: Submodular maximization over continuous domains," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 111–120.

[16] L. Tong, Y. Li, and W. Gao, "A hierarchical edge cloud architecture for mobile computing," in *IEEE INFOCOM 2016-IEEE International Conference on Computer Communications*, 2016, pp. 1–9.

[17] K. Poularakis, J. Llorca, A. M. Tulino, I. Taylor, and L. Tassiulas, "Joint service placement and request routing in multi-cell mobile edge computing networks," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, 2019, pp. 10–18.

[18] K. Kamran, E. Yeh, and Q. Ma, "Deco: Joint computation, caching and forwarding in data-centric computing networks," in *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2019, pp. 111–120.
[19] A. Lalitha, O. C. Kilinc, T. Javidi, and F. Koushanfar, "Peer-to-peer federated learning on graphs," *arXiv preprint arXiv:1901.11173*, 2019.
[20] G. Neglia, G. Calbi, D. Towsley, and G. Vardoyan, "The role of network topology for distributed machine learning," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, 2019, pp. 2350–2358.
[21] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, 2018, pp. 63–71.
[22] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5872–5881.
[23] S. Wang, Y. Ruan, Y. Tu, S. Wagle, C. G. Brinton, and C. Joe-Wong, "Network-aware optimization of distributed learning for fog computing," *IEEE/ACM Transactions on Networking*, 2021.
[24] F. Pukelsheim, *Optimal design of experiments*. Society for Industrial and Applied Mathematics, 2006.
[25] X. Huan and Y. M. Marzouk, "Simulation-based optimal bayesian experimental design for nonlinear systems," *Journal of Computational Physics*, vol. 232, no. 1, pp. 288–317, 2013.
[26] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.
[27] G. Calinescu, C. Chekuri, M. Pal, and J. Vondrák, "Maximizing a monotone submodular function subject to a matroid constraint," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1740–1766, 2011.
[28] T. Soma and Y. Yoshida, "A generalization of submodular cover via the diminishing return property on the integer lattice," *Advances in Neural Information Processing Systems*, vol. 28, pp. 847–855, 2015.
[29] H. Hassani, M. Soltanolkotabi, and A. Karbasi, "Gradient methods for submodular maximization," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 5843–5853.
[30] A. A. Ageev and M. I. Sviridenko, "Pipage rounding: A new method of constructing algorithms with proven performance guarantee," *Journal of Combinatorial Optimization*, vol. 8, no. 3, pp. 307–328, 2004.
[31] C. Chekuri, J. Vondrák, and R. Zenklusen, "Dependent randomized rounding via exchange properties of combinatorial structures," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 575–584.
[32] T. Soma and Y. Yoshida, "Maximizing monotone submodular functions over the integer lattice," *Mathematical Programming*, vol. 172, no. 1, pp. 539–563, 2018.
[33] S. Ioannidis and E. Yeh, "Adaptive caching networks with optimality guarantees," *IEEE/ACM Transactions on Networking*, vol. 26, no. 2, pp. 737–750, 2018.
[34] K. Poularakis and L. Tassiulas, "On the complexity of optimal content placement in hierarchical caching networks," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 2092–2103, 2016.
[35] S. Ioannidis and E. Yeh, "Jointly optimal routing and caching for arbitrary network topologies," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1258–1275, 2018.
[36] K. Kamran, A. Moharrer, S. Ioannidis, and E. Yeh, "Rate allocation and content placement in cache networks," in *IEEE INFOCOM 2021–IEEE Conference on Computer Communications*, 2021.
[37] T. Wu, P. Yang, H. Dai, W. Xu, and M. Xu, "Charging oriented sensor placement and flexible scheduling in rechargeable wsns," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, 2019, pp. 73–81.
[38] G. Sallam and B. Ji, "Joint placement and allocation of virtual network functions with budget and capacity constraints," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, 2019, pp. 523–531.
[39] Z. Zheng and N. B. Shroff, "Submodular utility maximization for deadline constrained data collection in sensor networks," *IEEE Transactions on Automatic Control*, vol. 59, no. 9, pp. 2400–2412, 2014.
[40] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, 2012, pp. 173–184.
[41] A. Krause and C. Guestrin, "Beyond convexity: Submodularity in machine learning," *ICML Tutorials*, 2008.
[42] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.
[43] R. G. Gallager, *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013.
[44] F. P. Kelly, *Reversibility and stochastic networks*. Cambridge University Press, 2011.
[45] C. L. Canonne, "A short note on poisson tail bounds," http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf.
[46] N. Alon and J. H. Spencer, *The probabilistic method*. John Wiley & Sons, 2004.
[47] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.
[48] J. Kleinberg, "The small-world phenomenon: An algorithmic perspective," in *Proceedings of the thirty-second Annual ACM Symposium on Theory of Computing*, 2000, pp. 163–170.
[49] D. Rossi and G. Rossini, "Caching performance of content centric networks under multi-path routing (and more)," *Relatório técnico, Telecom ParisTech*, pp. 1–6, 2011.
[50] R. Srikant, *The Mathematics of Internet Congestion Control*. Birkhäuser Boston, MA: Springer Science & Business Media, 2012.
[51] T. S. Jaakkola and M. I. Jordan, "A variational approach to bayesian logistic regression models and their extensions," in *Sixth International Workshop on Artificial Intelligence and Statistics*. PMLR, 1997, pp. 283–294.
[52] K. B. Petersen and M. S. Pedersen, "The matrix cookbook (version: November 15, 2012)," 2012.
[53] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge university press, 2012.

APPENDIX A
PROOF OF LEMMA 1

Proof. We extend the proof in Appendix A of [10] for the supmodularity of D-optimal design over a set to the integer lattice: For $\boldsymbol{n} \in \mathbb{N}^p$ and $k \in \mathbb{N}$, we have

$$\begin{aligned} &G(\boldsymbol{n} + k\boldsymbol{e}_i) - G(\boldsymbol{n}) = \\ &\quad = \log\det(\mathbf{I}_d + \frac{k}{\sigma^2}\boldsymbol{x}_i\boldsymbol{x}_i^\top(\Sigma_0^{-1} + \sum_{i=1}^{p}\frac{n_i}{\sigma^2}\boldsymbol{x}_i\boldsymbol{x}_i^\top)^{-1}) \\ &\quad = \log(1 + \frac{k}{\sigma^2}\boldsymbol{x}_i^\top A(\boldsymbol{n})^{-1}\boldsymbol{x}_i), \end{aligned}$$

where $A(\boldsymbol{n}) = \Sigma_0^{-1} + \sum_{i=1}^{p}\frac{n_i}{\sigma^2}\boldsymbol{x}_i\boldsymbol{x}_i^\top$, and the last equality follows Eq. (24) in [52]. The monotonicity of G follows because $A(\boldsymbol{n})^{-1}$ is positive semidefinite. Finally, since the matrix inverse is decreasing over the positive semi-definite order, we have $A(\boldsymbol{n})^{-1} \succeq A(\boldsymbol{m})^{-1}, \forall\, \boldsymbol{n}, \boldsymbol{m} \in \mathbb{N}^p$ and $\boldsymbol{n} \leq \boldsymbol{m}$, which leads to $G(\boldsymbol{n}+k\boldsymbol{e}_i) - G(\boldsymbol{n}) \geq G(\boldsymbol{m}+k\boldsymbol{e}_i) - G(\boldsymbol{m})$. □

APPENDIX B
PROOF OF LEMMA 2

Proof. To start with, $\boldsymbol{\lambda}^K \in \mathcal{D}$, as a convex combination of points in $\mathcal{D}$. Next, consider the point $\boldsymbol{v}^* = (\boldsymbol{\lambda}^* \vee \boldsymbol{\lambda}) - \boldsymbol{\lambda} = (\boldsymbol{\lambda}^* - \boldsymbol{\lambda}) \vee \mathbf{0} \geq \mathbf{0}$, in which $\boldsymbol{\lambda}$ is the solution at current iteration and $\boldsymbol{\lambda}^* \in \mathcal{D}$ is the optimal solution. Because U is non-decreasing (Thm. 1), we have

$$U(\boldsymbol{\lambda} + \boldsymbol{v}^*) = U(\boldsymbol{\lambda}^* \vee \boldsymbol{\lambda}) \geq U(\boldsymbol{\lambda}^*). \tag{35}$$

A DR-submodular continuous function is concave along any non-negative direction, and any non-positive direction (see

e.g., Prop. 4 in [15]), thus $g(\xi) := U(\boldsymbol{\lambda}+\xi\boldsymbol{v}^*)$, where $g'(\xi) = \langle \boldsymbol{v}^*, \nabla U(\boldsymbol{\lambda}+\xi\boldsymbol{v}^*)\rangle$, is concave, hence,

$$U(\boldsymbol{\lambda}^*) - U(\boldsymbol{\lambda}) = g(1) - g(0) \leq g'(0) \times 1 = \langle \boldsymbol{v}^*, \nabla U(\boldsymbol{\lambda})\rangle. \tag{36}$$

Then,

$$\begin{aligned}
&\langle \boldsymbol{v}, \nabla U(\boldsymbol{\lambda})\rangle \overset{(26)}{\geq} a \max_{\boldsymbol{v}\in\mathcal{D}} \langle \boldsymbol{v}, \nabla U(\boldsymbol{\lambda})\rangle - b \geq a\langle \boldsymbol{\lambda}^*, \nabla U(\boldsymbol{\lambda})\rangle - b \\
&\geq a\langle \boldsymbol{v}^*, \nabla U(\boldsymbol{\lambda})\rangle - b \overset{(36)}{\geq} a(U(\boldsymbol{\lambda}+\boldsymbol{v}^*) - U(\boldsymbol{\lambda})) - b \\
&\overset{(35)}{\geq} a(U(\boldsymbol{\lambda}^*) - U(\boldsymbol{\lambda})) - b,
\end{aligned} \tag{37}$$

where the second inequality is because the LHS maximizes the inner product, and the third inequality is because $\mathbf{0} \leq \boldsymbol{v}^* \leq \boldsymbol{\lambda}^*$ and $\nabla U(\boldsymbol{\lambda})$ is positive (Thm. 1). From the definition of the Hessian from Eqs. (24) and (25), we can show that $\|\nabla^2 U\|_F$ is upper bounded by $2|\mathcal{X}||\mathcal{L}|T^2 G_{\text{MAX}}$. The largest eigenvalue is upper bounded by Lipschitz continuous constant L, and $\max|\lambda(\nabla^2 U)| = \|\nabla^2 U\|_2 \leq \|\nabla^2 U\|_F$ [53]. Thus $L = 2|\mathcal{X}||\mathcal{L}|T^2 G_{\text{MAX}}$ is the Lipschitz continuous constant of ∇U . Then, we have

$$\begin{aligned}
&U(\boldsymbol{\lambda}^{k+1}) - U(\boldsymbol{\lambda}^k) = U(\boldsymbol{\lambda}^k + \gamma_k \boldsymbol{v}^k) - U(\boldsymbol{\lambda}^k) \\
&\geq \gamma_k \langle \boldsymbol{v}^k, \nabla U(\boldsymbol{\lambda}^k)\rangle - \frac{L}{2}\gamma_k^2 \|\boldsymbol{v}^k\|_2^2 \text{ (Lipschitz)} \\
&\overset{(37)}{\geq} a\gamma_k(\max_{\boldsymbol{\lambda}\in\mathcal{D}} U(\boldsymbol{\lambda}) - U(\boldsymbol{\lambda}^k)) - \gamma_k b - \frac{L}{2}\gamma_k^2 \|\boldsymbol{v}^k\|_2^2
\end{aligned}$$

After rearrangement, we have

$$\begin{aligned}
&U(\boldsymbol{\lambda}^{k+1}) - \max_{\boldsymbol{\lambda}\in\mathcal{D}} U(\boldsymbol{\lambda}) \geq \\
&(1 - a\gamma_k)[U(\boldsymbol{\lambda}^k) - \max_{\boldsymbol{\lambda}\in\mathcal{D}} U(\boldsymbol{\lambda})] - \gamma_k b - \frac{L}{2}\gamma_k^2 \lambda_{\text{MAX}}^2,
\end{aligned}$$

since $\|\boldsymbol{v}^k\|_2^2 \leq \lambda_{\text{MAX}}^2$. By telescope,

$$\begin{aligned}
&U(\boldsymbol{\lambda}^K) - \max_{\boldsymbol{\lambda}\in\mathcal{D}} U(\boldsymbol{\lambda}) \geq \\
&[U(\boldsymbol{\lambda}^0) - \max_{\boldsymbol{\lambda}\in\mathcal{D}} U(\boldsymbol{\lambda})]e^{-a} - b - \frac{L}{2}\sum_{k=0}^{K-1} \gamma_k^2 \lambda_{\text{MAX}}^2.
\end{aligned}$$

Finally, as $U(\boldsymbol{\lambda}^0) = 0$ and $\gamma_k = \delta = 1/K$, we have

$$U(\boldsymbol{\lambda}^K) \geq (1 - e^{-a})U(\boldsymbol{\lambda}^*) - \frac{L}{2}\delta\lambda_{\text{MAX}}^2 - b.$$

□

APPENDIX C
PROOF OF LEMMA 4

Proof. We further define

$$\text{TAIL}_{\boldsymbol{x}}^{\ell} = \sum_{n=n'+1}^{\infty} \Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, n) \cdot T \cdot \Pr[n_{\boldsymbol{x}}^{\ell} = n].$$

We have

$$\text{HEAD}_{\boldsymbol{x}}^{\ell} \geq \Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, n') \cdot T \cdot \Pr[n_{\boldsymbol{x}}^{\ell} \leq n']$$

and

$$\text{TAIL}_{\boldsymbol{x}}^{\ell} \leq \Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, n') \cdot T \cdot \Pr[n_{\boldsymbol{x}}^{\ell} \geq n'+1],$$

since $\Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, t_1) \geq \Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, t_2)$ for $t_1 \leq t_2$, resulting from the submodularity of G (Lemma 1). We note that $\frac{\partial U}{\partial \lambda_{\boldsymbol{x}}^{\ell}} = \text{HEAD}_{\boldsymbol{x}}^{\ell} + \text{TAIL}_{\boldsymbol{x}}^{\ell}$. Then we have,

$$\begin{aligned}
&\frac{\text{HEAD}_{\boldsymbol{x}}^{\ell}}{\text{HEAD}_{\boldsymbol{x}}^{\ell} + \text{TAIL}_{\boldsymbol{x}}^{\ell}} = \frac{1}{1 + \frac{\text{TAIL}_{\boldsymbol{x}}^{\ell}}{\text{HEAD}_{\boldsymbol{x}}^{\ell}}} \geq \\
&\frac{1}{1 + \frac{\Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, n') \cdot T \cdot \Pr[n_{\boldsymbol{x}}^{\ell} \geq n'+1]}{\Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, n') \cdot T \cdot (1 - \Pr[n_{\boldsymbol{x}}^{\ell} \geq n'+1])}} = 1 - \Pr[n_{\boldsymbol{x}}^{\ell} \geq n'+1],
\end{aligned}$$

thus,

$$\begin{aligned}
\text{HEAD}_{\boldsymbol{x}}^{\ell} &\geq (1 - \Pr[n_{\boldsymbol{x}}^{\ell} \geq n'+1]) \frac{\partial U}{\partial \lambda_{\boldsymbol{x}}^{\ell}} \\
&\geq (1 - e^{-\frac{(n' - \lambda_{\boldsymbol{x}}^{\ell} T + 1)^2}{2\lambda_{\boldsymbol{x}}^{\ell} T} h(\frac{n' - \lambda_{\boldsymbol{x}}^{\ell} T + 1}{\lambda_{\boldsymbol{x}}^{\ell} T})}) \frac{\partial U}{\partial \lambda_{\boldsymbol{x}}^{\ell}},
\end{aligned}$$

where $h(u) = 2\frac{(1+u)\ln(1+u) - u}{u^2}$ (Lemma 3). □

APPENDIX D
PROOF OF LEMMA 5

Proof. Our final estimator of the partial derivative is given by

$$\widehat{\frac{\partial U}{\partial \lambda_{\boldsymbol{x}}^{\ell}}} \equiv T \sum_{n=0}^{n'} \widehat{\Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, n)} \Pr[n_{\boldsymbol{x}}^{\ell} = n],$$

where

$$\widehat{\Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, n)} = \frac{1}{N}\sum_{j=1}^{N} (G^{\ell}(\boldsymbol{n}^{\ell,j}|_{n_{\boldsymbol{x}}^{\ell,j} = n+1}) - G^{\ell}(\boldsymbol{n}^{\ell,j}|_{n_{\boldsymbol{x}}^{\ell,j} = n})).$$

We define

$$X^j(n) = \frac{G^{\ell}(\boldsymbol{n}^{\ell,j}|_{n_{\boldsymbol{x}}^{\ell,j}=n+1}) - G^{\ell}(\boldsymbol{n}^{\ell,j}|_{n_{\boldsymbol{x}}^{\ell,j}=n}) - \Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, n)}{G_{\text{MAX}}},$$

where

$$G_{\text{MAX}} = \max_{\ell\in\mathcal{L}, \boldsymbol{x}\in\mathcal{X}} (G^{\ell}(\boldsymbol{e}_{\boldsymbol{x}}) - G^{\ell}(\mathbf{0})).$$

We have $|X^j(n)| \leq 1$, because

$$\begin{aligned}
G_{\text{MAX}} &\geq G^{\ell}(\boldsymbol{e}_{\boldsymbol{x}}) - G^{\ell}(\mathbf{0}) \\
&\geq G^{\ell}(\boldsymbol{n}^{\ell} | n_{\boldsymbol{x}} = n+1) - G^{\ell}(\boldsymbol{n}^{\ell} | n_{\boldsymbol{x}} = n),
\end{aligned}$$

for any $\ell \in \mathcal{L}, \boldsymbol{x} \in \mathcal{X}$, $n \geq 0$. By Chernoff bounds described by Theorem A.1.16 in [46], we have

$$\Pr\left[\left|\sum_{j=1}^{N}\sum_{n=0}^{n=n'} X^j(n)\right| > c\right] \leq 2e^{-c^2/2N(n'+1)}.$$

Suppose we let $c = \delta \cdot N/T$, where δ is the step size, then we have

$$\begin{aligned}
&\left|\widehat{\frac{\partial U}{\partial \lambda_{\boldsymbol{x}}^{\ell}}} - \text{HEAD}_{\boldsymbol{x}}^{\ell}\right| \\
&\leq \left|\sum_{n=0}^{n'}\sum_{j=1}^{N} \frac{G^{\ell}(\boldsymbol{n}^{\ell,j}|_{n_{\boldsymbol{x}}^{\ell,j}=n+1}) - G^{\ell}(\boldsymbol{n}^{\ell,j}|_{n_{\boldsymbol{x}}^{\ell,j}=n}) - \Delta_{\boldsymbol{x}}^{\ell}(\boldsymbol{\lambda}^{\ell}, n)}{N} T\right| \\
&= \left|\sum_{t=0}^{n'}\sum_{j=1}^{N} X^j(n)\right| \cdot \frac{T}{N} \cdot G_{\text{MAX}} \leq \delta \cdot G_{\text{MAX}},
\end{aligned}$$

with probability greater than $1 - 2 \cdot e^{-\delta^2 N/2T^2(n'+1)}$. By Lemma 4, we have $\frac{\partial U}{\partial \lambda^\ell_{\boldsymbol{x}}} \geq \text{HEAD}^\ell_{\boldsymbol{x}} \geq (1 - \text{P}[n^\ell_{\boldsymbol{x}} \geq n'+1]) \cdot \frac{\partial U}{\partial \lambda^\ell_{\boldsymbol{x}}}$. Thus, we have

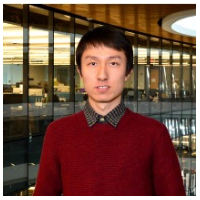
$$-\delta \cdot G_{\text{MAX}} \leq \frac{\partial U}{\partial \lambda^\ell_{\boldsymbol{x}}} - \widehat{\frac{\partial U}{\partial \lambda^\ell_{\boldsymbol{x}}}} \leq \delta \cdot G_{\text{MAX}} + \Pr[n^\ell_{\boldsymbol{x}} \geq n'+1] \cdot \frac{\partial U}{\partial \lambda^\ell_{\boldsymbol{x}}}. \tag{38}$$

We now use the superscript k to represent the parameters for the kth iteration: we find $\boldsymbol{v}^k \in \mathcal{D}$ that maximizes $\langle \boldsymbol{v}^k, \widehat{\nabla U(\boldsymbol{\lambda}^k)} \rangle$. Let $\boldsymbol{u}^k \in \mathcal{D}$ be the vector that maximizes $\langle \boldsymbol{u}^k, \nabla U(\boldsymbol{\lambda}^k) \rangle$ instead and define $\text{P}_{\text{MAX}} = \max_{k=1,\ldots,K} \text{P}^k_{\text{MAX}}$ where $\text{P}^k_{\text{MAX}} = \max_{l \in \mathcal{L}, \boldsymbol{x} \in \mathcal{X}} \text{P}[n^{\ell,k}_{\boldsymbol{x}} \geq t'+1]$ and $\lambda_{\text{MAX}} \equiv \max_{\boldsymbol{\lambda} \in \mathcal{D}} \sum_{\ell \in \mathcal{L}} \|\boldsymbol{\lambda}^\ell\|_1$. We have

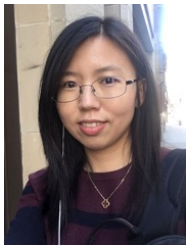
$$\begin{aligned}
&\langle \boldsymbol{v}^k, \nabla U(\boldsymbol{\lambda}^k) \rangle \geq \langle \boldsymbol{v}^k, \widehat{\nabla U(\boldsymbol{\lambda}^k)} \rangle - \lambda_{\text{MAX}} \delta \cdot G_{\text{MAX}} \\
\geq &\langle \boldsymbol{u}^k, \widehat{\nabla U(\boldsymbol{\lambda}^k)} \rangle - \lambda_{\text{MAX}} \delta \cdot G_{\text{MAX}} \\
\geq &(1 - \text{P}_{\text{MAX}}) \cdot \langle \boldsymbol{u}^k, \nabla U(\boldsymbol{\lambda}^k) \rangle - 2\lambda_{\text{MAX}} \delta \cdot G_{\text{MAX}},
\end{aligned}$$

where the first and last inequalities are due to (38) and the second inequality is because $\boldsymbol{v}^k$ maximizes $\langle \boldsymbol{v}^k, \widehat{\nabla U(\boldsymbol{\lambda}^k)} \rangle$. The above inequality requires the satisfaction of (38) for every partial derivative. By union bound, the above inequality satisfies with probability greater than $1 - 2|\mathcal{X}\|\mathcal{L}| \cdot e^{-\delta^2 N/2T^2(n'+1)}$. □

Yuanyuan Li received the B.E. degree (2015) from School of Electronic and Information Engineering, South China University of Technology and M.S. degree (2018) from Department of Computer Science and Engineering, Shanghai Jiao Tong University. She is now a Ph.D. candidate in Computer Engineering, Northeastern University, Boston, USA., under the supervision of Prof. Stratis Ioannidis. Her research interests include networking, optimization and machine learning.



Yuezhou Liu received his B.E. (2018) in Information Engineering from Shanghai Jiao Tong University. He received his M.S. (2021) and is now a Ph.D. student in Electrical Engineering at Northeastern University, under the supervision of Prof. Edmund Yeh. His research interests include networking, edge computing, and distributed machine learning.



Lili Su is an Assistant Professor in the Electrical and Computer Engineering department at Northeastern University, in Boston, MA. She also holds a courtesy appointment with the Khoury College of Computer Sciences. She received her M.Sc. (2014) and Ph.D. (2017) in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign (UIUC). Prior to joining Northeastern, she was a postdoc in the Computer Science and Articial Intelligence Laboratory (CSAIL) at MIT. Her research intersects distributed systems - resilience and efficiency, distributed learning, federated learning, and multi-agent systems. She is the runner-up of the Best Student Paper Award from the 30th International Symposium on Distributed Computing (DISC 2016), and she received the 2015 Best Student Paper Award from the 17th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2015). She received Sundaram Seshu International Student Fellowship from UIUC in 2016, and was selected as Rising Stars in EECS (2018).

Edmund Yeh received the B.S. degree (Hons.) and Phi Beta Kappa in electrical engineering from Stanford University in 1994, the M.Phil. degree in engineering from Cambridge University on the Winston Churchill Scholarship in 1995, and the Ph.D. degree in electrical engineering and computer science from MIT under Prof. Robert Gallager in 2001. He is currently a Professor of Electrical and Computer Engineering with Northeastern University, with a courtesy appointment at the Khoury School of Computer Sciences. He was previously an Assistant and an Associate Professor of Electrical Engineering, Computer Science, and Statistics with Yale University. He is an IEEE Communications Society Distinguished Lecturer. He was a recipient of the Alexander von Humboldt Research Fellowship, the Army Research Office Young Investigator Award, the Winston Churchill Scholarship, the National Science Foundation and Office of Naval Research Graduate Fellowships, the Barry M. Goldwater Scholarship, the Frederick Emmons Terman Engineering Scholastic Award, and the President's Award for Academic Excellence (Stanford University). He has received three best paper awards, including awards from the 2017 ACM Conference on Information-Centric Networking (ICN), and the 2015 IEEE International Conference on Communications (ICC) Communication Theory Symposium. He serves as a TPC Co-Chair for ACM MobiHoc 2021. He also serves as a Treasurer of the Board of Governors for the IEEE Information Theory Society. He served as the General Chair for ACM SIGMETRICS 2020, an Associate Editor for IEEE TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON MOBILE COMPUTING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, as a Guest Editor-in-Chief of the Special Issue on Wireless Networks for Internet Mathematics, and a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS–Special Series on Smart Grid Communications.

Stratis Ioannidis is an Associate Professor in the Electrical and Computer Engineering department at Northeastern University, in Boston, MA, where he also holds a courtesy appointment with the Khoury College of Computer Sciences. He received his B.Sc. (2002) in Electrical and Computer Engineering from the National Technical University of Athens, Greece, and his M.Sc. (2004) and Ph.D. (2009) in Computer Science from the University of Toronto, Canada. Prior to joining Northeastern, he was a research scientist at the Technicolor research centers in Paris, France, and Palo Alto, CA, as well as at Yahoo Labs in Sunnyvale, CA. He is the recipient of an NSF CAREER award, a Google Faculty Research Award, a Facebook Research Award, and Best Paper Awards at the 2017 ACM Conference on Information-centric Networking (ICN) and the 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN).