Decentralized Prediction-Correction Methods for Networked Time-Varying Convex Optimization

Andrea Simonetto, *Member, IEEE*, Alec Koppel, Aryan Mokhtari, Geert Leus, *Fellow, IEEE*, and Alejandro Ribeiro, *Member, IEEE*

***Abstract*—We develop algorithms that find and track the optimal solution trajectory of time-varying convex optimization problems that consist of local and network-related objectives. The algorithms are derived from the prediction-correction methodology, which corresponds to a strategy where the time-varying problem is sampled at discrete time instances, and then, a sequence is generated via alternatively executing predictions on how the optimizers at the next time sample are changing and corrections on how they actually have changed. Prediction is based on how the optimality conditions evolve in time, while correction is based on a gradient or Newton method, leading to decentralized prediction-correction gradient and decentralized prediction-correction Newton. We extend these methods to cases where the knowledge on how the optimization programs are changing in time is only approximate and propose decentralized approximate prediction-correction gradient and decentralized approximate prediction-correction Newton. Convergence properties of all the proposed methods are studied and empirical performance is shown on an application of a resource allocation problem in a wireless network. We observe that the proposed methods outperform existing running algorithms by orders of magnitude. The numerical results showcase a tradeoff between convergence accuracy, sampling period, and network communications.**

***Index Terms*—Distributed algorithms, optimization methods, time-varying optimization, wireless sensor networks.**

I. Introduction

DECENTRALIZED tracking methods are used to solve problems in which distinct agents of a network aim at minimizing a global objective that varies continuously in time. We focus on a special case of this problem, where the objective may be decomposed into two parts: The first part is a sum of functions that are locally available at each node; the second is defined along the edges of the network, and is often defined by the cost of communication among the agents. Problems of this kind arise, e.g., in estimation, control, and robotics [3]–[9].

One approach to continuous-time optimization problems of this kind is to sample the objective function at discrete time instances t_k, $k = 0, 1, 2, \ldots$, and then, solve each time-invariant instance of the problem, via classical methods such as gradient or Newton descent. If the sampling period $h := t_{k+1} - t_k$ is chosen arbitrarily small, then doing so would yield the solution trajectory $\boldsymbol{y}^*(t_k)$ with arbitrary accuracy. However, solving such problems for each time sample is not a viable option in most application domains, since the computation time to obtain each optimizer exceeds the rate at which the solution trajectory changes, unless $\boldsymbol{y}^*(t)$ is approximately stationary.

Prediction-correction algorithms [10], by making use of tools of nonstationary optimization [11]–[13], have been developed to iteratively solve convex programs that continuously vary in time. These methods operate by predicting at time t_k the optimal solution at the discrete time instance t_{k+1} via an approximation of the variation of the objective function F over this time slot. Then, this prediction is revised by executing gradient or Newton descent. However, these methods are designed only for centralized settings. We focus on time-varying convex programs in decentralized settings, where nodes can only communicate with their neighbors. As a consequence, the prediction-correction methods suggested in [10] are not directly applicable.

One approach to solving problems of this type are decentralized running algorithms, which run at the same time scale as the optimization problem and dynamically react to changes in the objective function. Performance guarantees for such methods yield convergence to a neighborhood of the true optimizer $\boldsymbol{y}^*(t_k)$ on the order of the sampling period $O(h)$, despite the fact that only one round of communication is allowed per discrete time step [4], [14]–[20]. The aforementioned works mostly consider strongly convex objectives with no constraints. Notably, [18] and [19] describe a running dual decomposition and a running alternating direction method of multipliers algorithm. Notice that these methods implement only correction steps, and thus, cannot effectively mitigate the error from the non-stationarity of the optimizer.

In this paper, we generalize the prediction-correction methodology of [10] to decentralized settings such that each node of a

Manuscript received November 7, 2016; accepted March 28, 2017. Date of publication April 17, 2017; date of current version October 25, 2017. The work in this paper is supported by NSF CAREER CCF-0952867, ONR N00014-12-1-0997, ASEE SMART, and ARL MAST CTA. This paper expands the results and presents proofs that are referenced in [1], [2]. Recommended by Associate Editor C. Hadjicostis. *(Corresponding author: Andrea Simonetto.)*

A. Simonetto is with IBM Research Ireland, Dublin 15, Ireland (e-mail: andrea.simonetto@ibm.com).

G. Leus is with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: g.j.t.leus@tudelft.nl).

A. Koppel, A. Mokhtari, and Alejandro Ribeiro are with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: akoppel@seas.upenn.edu; aryanm@seas.upenn.edu; aribeiro@seas.upenn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2017.2694611

network, after communicating with its neighbors, estimates its local component of the optimal trajectory at the discrete time instance t_{k+1} from information regarding the objective at time t_k, and then, corrects this local prediction at time t_{k+1}, via additional communications within the network. To develop this generalization, in the prediction step, we truncate the Taylor series of the objective function's Hessian inverse. This approximation is necessary since the computation of the objective function's Hessian inverse, which is required for the prediction step, requires global communication. In the correction step, we use decentralized approximations of gradient descent and of Newton's method to correct the predicted solution by descending towards the optimal solution of the observed objective function. In addition, we consider cases in which the prediction of how the cost function changes in time is unavailable, and must be estimated. This time-derivative approximation is particularly useful in target tracking [21] or designing learning-based control strategies [22], [23].

The main contributions of this paper are the following.

1) We develop prediction-correction algorithms for a class of time-varying networked optimization problems, which can be implemented in a distributed fashion over a network of computing and communicating nodes. The correction term is either derived from a gradient method or from a (damped) Newton step.

2) In order to compute the prediction (and correction for Newton) direction, we employ a novel matrix splitting technique, for which the one developed in [24] and[25] is a special case (only valid for adjacency matrices). The novel methodology relies on the concept of block diagonal dominance.

3) We prove convergence of all the algorithms and characterize their convergence rate. For the case of the (damped) Newton correction step, we compute the (local) convergence region and argue global convergence in case of a damped step.

This paper is organized as follows. In Section II, we begin by introducing the optimization problem of interest and by providing some examples for the proposed formulation. We then derive a family of algorithms, which contains four distinct methods (see Section III). We analyze their convergence properties in Section IV, establishing that the sequence of iterates generated by all these algorithms converges linearly to a bounded tracking error. We observe a tradeoff in the implementation between approximation accuracy and communication cost. In Section V, we numerically analyze the methods on a resource allocation problem in wireless sensor networks. Finally, in Section VI, we conclude.[1]

[1]*Notation:* Vectors are written as $\boldsymbol{y} \in \mathbb{R}^n$ and matrices as $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. $\|\cdot\|$ denotes the Euclidean norm, in the case of vectors, matrices, and tensors. The gradient of the function $f(\boldsymbol{y};t)$ with respect to $\boldsymbol{y}$ at the point $(\boldsymbol{y},t)$ is indicated as $\nabla_{\boldsymbol{y}} f(\boldsymbol{y};t) \in \mathbb{R}^n$, while the partial derivative of the same function w.r.t. t at $(\boldsymbol{y},t)$ is $\nabla_t f(\boldsymbol{y};t) \in \mathbb{R}$. Similarly, the notation $\nabla_{\boldsymbol{y}\boldsymbol{y}} f(\boldsymbol{y};t) \in \mathbb{R}^{n\times n}$ denotes the Hessian of $f(\boldsymbol{y};t)$ w.r.t. $\boldsymbol{y}$ at $(\boldsymbol{y},t)$, whereas $\nabla_{t\boldsymbol{y}} f(\boldsymbol{y};t) \in \mathbb{R}^n$ denotes the partial derivative of the gradient of $f(\boldsymbol{y};t)$, w.r.t., time t at $(\boldsymbol{y},t)$, i.e. the mixed first-order partial derivative vector of the objective. Consistent notation is used for higher order derivatives.

II. Problem Formulation

We consider a connected undirected graph $\mathcal{G} = (V, E)$, with vertex set V containing n nodes and edge set E containing m edges. Consider $\boldsymbol{y}^i \in \mathbb{R}^p$ as the decision variable of node i and t as a nonnegative scalar that represents time. Associated with each node i are time-varying strongly convex functions $f^i(\boldsymbol{y}^i;t) : \mathbb{R}^p \times \mathbb{R}_+ \to \mathbb{R}$ and $g^{i,i}(\boldsymbol{y}^i;t) : \mathbb{R}^p \times \mathbb{R}_+ \to \mathbb{R}$. The local functions f^i may be interpreted as, e.g., the merit of a particular choice of control policy [5] or statistical model [3]. Moreover, associated with each edge $(i,j) \in E$ is a continuously time-varying convex function $g^{i,j}(\boldsymbol{y}^i, \boldsymbol{y}^j;t) : \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}_+ \to \mathbb{R}$. These edge-wise functions represent, e.g., the cost of communicating across the network [26].

We focus on problems where nodes aim at cooperatively minimizing the global *smooth strongly convex* cost function $F : \mathbb{R}^{np} \times \mathbb{R}_+ \to \mathbb{R}$, which can be written as the sum of locally available functions $f : \mathbb{R}^{np} \times \mathbb{R}_+ \to \mathbb{R}$, and a function $g : \mathbb{R}^{np} \times \mathbb{R}_+ \to \mathbb{R}$ induced by the network structure $\mathcal{G}$. In particular, the function $f(\boldsymbol{y};t)$ is the sum of the locally available functions $f^i(\boldsymbol{y}^i;t)$, as follows:

$$f(\boldsymbol{y};t) := \sum_{i\in V} f^i(\boldsymbol{y}^i;t) \tag{1}$$

where we have defined $\boldsymbol{y} \in \mathbb{R}^{np}$ in (1) as the stacking of the nodes' decision variables $\boldsymbol{y}^i$, i.e., $\boldsymbol{y} = ({\boldsymbol{y}^1}^\mathsf{T}; \ldots; {\boldsymbol{y}^n}^\mathsf{T})^\mathsf{T}$. The function $g(\boldsymbol{y};t)$ induced by the structure of the network is the sum of locally available functions $g^{i,i}(\boldsymbol{y}^i;t)$ and the functions $g^{i,j}(\boldsymbol{y}^i, \boldsymbol{y}^j;t)$ associated to the edges of the network

$$g(\boldsymbol{y};t) := \sum_{i\in V} g^{i,i}(\boldsymbol{y}^i;t) + \sum_{(i,j)\in E} g^{i,j}(\boldsymbol{y}^i, \boldsymbol{y}^j;t)\,. \tag{2}$$

Our goal is to solve the time-varying convex program

$$\boldsymbol{y}^*(t) := \operatorname*{argmin}_{\boldsymbol{y}\in\mathbb{R}^{np}} F(\boldsymbol{y};t) := f(\boldsymbol{y};t) + g(\boldsymbol{y};t), \text{ for } t \geq 0 \tag{3}$$

that is the foundation of many problems in cooperative control and network utility maximization. Our goal is to enable the nodes to determine their own component of the solution $\boldsymbol{y}^*(t)$ of (3) for each time t in a decentralized fashion, i.e., a protocol such that each node only requires communication with neighboring nodes. Notice that nodes can minimize the objective function $f(\boldsymbol{y};t)$ independently, while minimization of the function $g(\boldsymbol{y};t)$ requires coordination and information exchange across the network. Before developing distributed protocols to solve (3), we present a couple of examples to clarify the problem setting.

Example 1 (Estimation of distributed processes): We consider a network of interconnected sensors monitoring a time-varying distributed process. We represent this process by a vector-valued function $\boldsymbol{u}(\boldsymbol{x},t) \in \mathbb{R}^p$, with $\boldsymbol{x} \in \mathbb{R}^3$ being the spatial coordinate, and t denoting time. We assume that the process is spatially smooth so that the value of $\boldsymbol{u}(\boldsymbol{x},t)$ at close-by spatial coordinates is also similar. We focus on the case that a network of n sensors is deployed in a spatial region $\mathcal{A} \subset \mathbb{R}^3$. The ith node acquires measurements $z^i(\boldsymbol{x}^i,t)$ that are noisy linear transformations of the true process $z^i(\boldsymbol{x}^i,t) = {\boldsymbol{h}^i}^\mathsf{T}\boldsymbol{u}(\boldsymbol{x}^i,t) + \eta^i(t)$, where $\boldsymbol{x}^i$ is the location of the sensor i, $\boldsymbol{h}^i$ is its regressor, and the noise $\eta^i(t) \sim \mathcal{N}(0,\sigma^i)$ is Gaussian distributed inde-

pendently across time with covariance σ^i. This problem setting comes up in earth sciences [27], [28] and acoustics [29], but it is also relevant in robotics [9], [30], [31]. By considering the task of learning a spatially regularized least-squares estimate $\hat{\boldsymbol{u}} \in \mathbb{R}^{np}$ of the process $\boldsymbol{u}(\boldsymbol{x}, t)$ at different locations, we obtain the time-varying networked convex program

$$\min_{\hat{\boldsymbol{u}}^1 \in \mathbb{R}^p, \ldots, \hat{\boldsymbol{u}}^n \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{h}^{i\,\mathsf{T}} \hat{\boldsymbol{u}}^i - z_i(\boldsymbol{x}^i, t)\|^2_{\frac{1}{\sigma^i}} + \frac{\beta}{2} \sum_{j \in N^i} w^{ij} \|\hat{\boldsymbol{u}}^i - \hat{\boldsymbol{u}}^j\|_2^2 \qquad (4)$$

where N^i denotes the neighborhood of node i, $\hat{\boldsymbol{u}}^i$ is the estimated value of the process $\boldsymbol{u}(\boldsymbol{x}, t)$ at time t and location $\boldsymbol{x}^i$, the constant $\beta > 0$ is a regularizer that incentivizes closely located sensors to obtain similar estimates, and the nonnegative weights w^{ij} may be defined according to a function of the distance between sensors. The first term in (4) defines the estimation accuracy in terms of the squared error and is identified as a sum of functions, which only depend on local information, which is a special case of (1). The second term in (4) couples the decisions of node i with its neighbors $j \in N^i$, and it is of the form (2). Thus, (4) is an instance of (3).

Example 2 (Resource allocation problems): Consider a resource allocation problem in a wireless sensor network [26], [32], [33]. Associate with sensor i a time-varying utility functions $f^i : \mathbb{R}^p \times \mathbb{R}_+$ and decision variable $\boldsymbol{y}^i \in \mathbb{R}^p$ representing the resources allocated to node i in a network $\mathcal{G}$ of n sensors. To allocate resources in this network, one must respect channel capacity and interference constraints. These constraints may be formulated in aggregate as network-flow constraints, obtaining the time-varying resource allocation problem

$$\min_{\boldsymbol{y}^1 \in \mathbb{R}^p, \ldots, \boldsymbol{y}^n \in \mathbb{R}^p} \sum_{i \in V} f^i(\boldsymbol{y}^i; t) \;\text{ subject to }\; \boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}(t)\,. \qquad (5)$$

In (5), $\boldsymbol{A} \in \mathbb{R}^{lp \times np}$ denotes the augmented graph edge incidence matrix. The matrix $\boldsymbol{A}$ is formed by $l \times n$ square blocks of dimension p. If the edge $e = (j, k)$ with $j < k$ links node j to node k the block (e, j) is $[\boldsymbol{A}]_{ej} = \boldsymbol{I}_p$ and the block $[\boldsymbol{A}]_{ek} = -\boldsymbol{I}_p$, where $\boldsymbol{I}_p$ denotes the identity matrix of dimension p. All other blocks are identically null. Moreover, the time-varying vectors $\boldsymbol{b}(t) \in \mathbb{R}^{lp}$ are induced by channel capacity and rate transmission constraints.

In many situations, especially in commercial settings where the nodes are consumer devices, one seeks to solve decentralized approximations of (5). One way to do so is to consider the approximate augmented Lagrangian relaxation of (5), and solve instead

$$\min_{\boldsymbol{y}^1 \in \mathbb{R}^p, \ldots, \boldsymbol{y}^n \in \mathbb{R}^p} \sum_{i \in V} f^i(\boldsymbol{y}^i; t) + \frac{1}{\beta^2} \|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}(t)\|^2 \qquad (6)$$

which is now unconstrained [34]. Notice that the parameter $\beta > 0$, which behaves similarly to a Lagrange multiplier, tunes the approximation level and penalizes the violation of the approximated constraint $\|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}(t)\|^2$. Observe that the first term in (6) is precisely the same as (1). Moreover, block-wise decomposition of the second term yields edge-wise expressions of the form $\|(\boldsymbol{y}_i - \boldsymbol{y}_j) - \boldsymbol{b}_i(t)\|^2$, which may be identified as the functions $g^{i,j}(\boldsymbol{y}^i, \boldsymbol{y}^j; t)$ in (2).

III. Algorithm Development

To solve the time-varying optimization problem in (3), the first step is sampling the continuously time-varying objective function $F(\boldsymbol{y}; t)$ at time instants t_k with $k = 0, 1, 2, \ldots$, leading to a sequence of time-invariant convex problems

$$\boldsymbol{y}^*(t_k) \;:=\; \operatorname*{argmin}_{\boldsymbol{y} \in \mathbb{R}^{np}} F(\boldsymbol{y}; t_k) \qquad k \geqslant 0\,. \qquad (7)$$

The sequence of optimal decision variables $\boldsymbol{y}^*(t_k)$ defined in (7) are samples of the optimal trajectory $\boldsymbol{y}^*(t)$ defined in (3). Since solving (7) for each time instance t_k is impractical even for moderately sized networks, we instead devise a method to generate a sequence of approximate optimizers for (7), which eventually remains close to the true optimizer $\boldsymbol{y}^*(t_k)$ in (7) up to a constant error. More formally, we seek to generate a sequence $\{\boldsymbol{y}_k\}$ for which

$$\limsup_{k \to \infty} \|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| = \text{const.} \qquad (8)$$

and whose rate, convergence, and asymptotical error constants depend on the sampling period h and the number of exchanged messages per node per time instance k.

To do so, we build upon prediction-correction methods, which at the current time sample t_k *predict* the optimal decision variable at the next time sample t_{k+1}, i.e., from an arbitrary initial variable $\boldsymbol{y}_0$, for each time $k \geqslant 0$, predict a new approximate optimizer as

$$\boldsymbol{y}_{k+1|k} = \boldsymbol{y}_k + h\,\boldsymbol{p}_k \qquad (9)$$

where index k is associated with time sample t_k, and similarly for $k+1$, w.r.t., t_{k+1}, $\boldsymbol{p}_k \in \mathbb{R}^{np}$ is the prediction direction, $\boldsymbol{y}_{k+1|k}$ is the predicted variable for step k+1, and h is the sampling period. Then, after observing the sampled objective function at t_{k+1}, we correct the predicted vector $\boldsymbol{y}_{k+1|k}$ by

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_{k+1|k} + \gamma\,\boldsymbol{c}_{k+1} \qquad (10)$$

for a certain correction direction $\boldsymbol{c}_{k+1} \in \mathbb{R}^{np}$, which defines a descent direction, with nonnegative constant step size $\gamma > 0$.

A. Decentralized Prediction Step

Solving the strongly convex time-invariant problem (7) accounts in finding the unique decision variable for which

$$\nabla_{\boldsymbol{y}} F(\boldsymbol{y}^*(t_k); t_k) = \boldsymbol{0}. \qquad (11)$$

For any other variable $\boldsymbol{y}_k \neq \boldsymbol{y}^*(t_k)$, the gradient $\nabla_{\boldsymbol{y}} F(\boldsymbol{y}_k; t_k)$ would not be null and we can use it to quantify the suboptimality of $\boldsymbol{y}$, w.r.t., $\boldsymbol{y}^*(t_k)$.

We design the prediction direction as the one that maintains the suboptimality level when determining $\boldsymbol{y}_{k+1|k}$ (the rationale being that when arrived at optimality, we will keep it while predicting). Formally, we wish to determine $\boldsymbol{y}_{k+1|k}$ as the vector for which

$$\nabla_{\boldsymbol{y}} F(\boldsymbol{y}_{k+1|k}; t_{k+1}) = \nabla_{\boldsymbol{y}} F(\boldsymbol{y}_k; t_k). \qquad (12)$$

Of course, implementing (12) requires information at future times t_{k+1} at the present t_k, an impossibility without clairvoyance. Instead, we approximate the left-hand side by adopting a Taylor expansion, obtaining

$$\nabla_{\boldsymbol{y}} F(\boldsymbol{y}_k;t_k) + \nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k)(\boldsymbol{y}_{k+1|k}-\boldsymbol{y}_k) + h\,\nabla_{t\boldsymbol{y}} F(\boldsymbol{y}_k;t_k) = \nabla_{\boldsymbol{y}} F(\boldsymbol{y}_k;t_k) \tag{13}$$

which may be reordered so that $\boldsymbol{y}_{k+1|k}$ is on the left-hand side, yielding

$$\boldsymbol{y}_{k+1|k} = \boldsymbol{y}_k - h\,[\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k)]^{-1}\nabla_{t\boldsymbol{y}} F(\boldsymbol{y}_k;t_k)\,. \tag{14}$$

The update (14) describes the discrete-time iso-suboptimality dynamics. This prediction step (14) in principle would allow us to maintain a consistent level of suboptimality, but our focus on decentralized methods precludes its use. This is because execution of (14) requires computing the Hessian inverse $\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k)^{-1}$, which is not implementable by a network due to the fact that $\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k)^{-1}$ is a global computation. The Hessian $\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k) = \nabla_{\boldsymbol{yy}} f(\boldsymbol{y};t) + \nabla_{\boldsymbol{yy}} g(\boldsymbol{y};t)$ consists of two terms: The first term $\nabla_{\boldsymbol{yy}} f(\boldsymbol{y};t)$ is a block diagonal matrix and the second term $\nabla_{\boldsymbol{yy}} g(\boldsymbol{y};t)$ is a block neighbor sparse matrix that inherits the structure of the graph. Therefore, the global objective function's Hessian $\nabla_{\boldsymbol{yy}} F(\boldsymbol{y};t)$ has the sparsity pattern of the graph and can be computed by exchanging information with neighboring nodes. Nonetheless, the Hessian inverse, required in (14), is not neighbor sparse and its computation requires global information.

To develop a decentralized protocol to approximately execute (14), we generalize a recently proposed technique to approximate the Hessian inverse $[\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k)]^{-1}$, which operates by truncating its Taylor expansion [24], [25]. To do so, define $\mathrm{diag}[\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_k;t_k)]$ as the block diagonal matrix, which contains the diagonal blocks of the matrix $\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_k;t_k)$, and write the Hessian $\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k)$ as

$$\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k) = \mathbf{D}_k - \mathbf{B}_k \tag{15}$$

where the matrices $\mathbf{D}_k$ and $\mathbf{B}_k$ are defined as

$$\mathbf{D}_k := \nabla_{\boldsymbol{yy}} f(\boldsymbol{y}_k;t_k) + \mathrm{diag}[\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_k;t_k)] \tag{16a}$$

$$\mathbf{B}_k := \mathrm{diag}[\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_k;t_k)] - \nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_k;t_k)\,. \tag{16b}$$

Since F is strongly convex, and by Assumption 2 [Cf., Section IV], the matrix $\mathbf{D}_k$ is a positive definite block diagonal matrix and encodes second-order local objective information. The structure of the matrix $\mathbf{B}_k$ is induced by that of the graph: The diagonal blocks of $\mathbf{B}_k$ are null and the nondiagonal block $\mathbf{B}_k^{ij}$ is nonzero and given by $-\nabla_{\boldsymbol{y}^i\boldsymbol{y}^j} g^{i,j}(\boldsymbol{y}_k^i,\boldsymbol{y}_k^j;t_k)$ iff i and j are neighbors.

Given that $\mathbf{D}_k$ is positive definite, we can write

$$\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k) = \mathbf{D}_k^{1/2}(\mathbf{I} - \mathbf{D}_k^{-1/2}\mathbf{B}_k\mathbf{D}_k^{-1/2})\mathbf{D}_k^{1/2}\,. \tag{17}$$

Consider now the Taylor series $(\mathbf{I}-\boldsymbol{X})^{-1} = \sum_{\tau=0}^{\infty}\boldsymbol{X}^{\tau}$ for $\boldsymbol{X} = \mathbf{D}_k^{-1/2}\mathbf{B}_k\mathbf{D}_k^{-1/2}$ to write the inverse of (17) as

$$[\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k)]^{-1} = \mathbf{D}_k^{-1/2}\sum_{\tau=0}^{\infty}\left(\mathbf{D}_k^{-1/2}\mathbf{B}_k\mathbf{D}_k^{-1/2}\right)^{\tau}\mathbf{D}_k^{-1/2} \tag{18}$$

whose convergence (as well as the fact that the absolute values of all the eigenvalues of X are strictly less than one so that the Taylor series holds) will be formally proved in Appendix A. We approximate the Hessian inverse $[\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k;t_k)]^{-1}$ in (18) by its Kth order approximate $\mathbf{H}_{k,(K)}^{-1}$, which is formed by truncating the series in (18) to its first $K+1$ terms as

$$\mathbf{H}_{k,(K)}^{-1} = \mathbf{D}_k^{-1/2}\sum_{\tau=0}^{K}\left(\mathbf{D}_k^{-1/2}\mathbf{B}_k\mathbf{D}_k^{-1/2}\right)^{\tau}\mathbf{D}_k^{-1/2}\,. \tag{19}$$

Since the matrix $\mathbf{D}_k$ is block diagonal and $\mathbf{B}_k$ is block neighbor sparse, it follows that the Kth order approximate inverse $\mathbf{H}_{k,(K)}^{-1}$ is K-hop block neighbor sparse, i.e., its ij-th block is nonzero if there is a path between nodes i and j with length K or smaller. Substituting the approximation in (19) into (14), the prediction step may be written as

$$\boldsymbol{y}_{k+1|k} = \boldsymbol{y}_k + h\,\boldsymbol{p}_{k,(K)} \tag{20}$$

where the approximate prediction direction $\boldsymbol{p}_{k,(K)}$ is given by

$$\boldsymbol{p}_{k,(K)} := -\mathbf{H}_{k,(K)}^{-1}\nabla_{t\boldsymbol{y}} F(\boldsymbol{y}_k;t_k). \tag{21}$$

Although the computation of the approximate prediction direction $\boldsymbol{p}_{k,(K)}$ requires information of K-hop neighbors, we establish that it can be computed in a decentralized manner via K communication rounds among neighboring nodes.

Proposition 1: Consider the prediction step (20) and the approximate prediction direction $\boldsymbol{p}_{k,(K)}$ in (21). Define $\boldsymbol{p}_{k,(K)}^i$ and $\nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k;t_k)$ as the ith subvector of the vectors $\boldsymbol{p}_{k,(K)}$ and $\nabla_{t\boldsymbol{y}} F(\boldsymbol{y}_k;t_k)$ associated with node i. Consider $\mathbf{D}_k^{ij}$ and $\mathbf{B}_k^{ij}$ as the ijth block of the matrices $\mathbf{D}_k$ and $\mathbf{B}_k$ in (16a)–(16b). If node i computes for $\tau = 0,\dots,K-1$, the recursion

$$\boldsymbol{p}_{k,(\tau+1)}^i = -(\mathbf{D}_k^{ii})^{-1}\left(\sum_{j\in N^i}\mathbf{B}_k^{ij}\boldsymbol{p}_{k,(\tau)}^j + \nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k;t_k)\right) \tag{22}$$

with initial condition $\boldsymbol{p}_{k,(0)}^i = -(\mathbf{D}_k^{ii})^{-1}\nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k;t_k)$, the result yields the approximate prediction direction $\boldsymbol{p}_{k,(K)}^i$.

Proof: By direct computation. See [24, Sec. III] for a comparable derivation. ■

The recursion in (22) allows for the computation the K-th order approximate prediction direction $\boldsymbol{p}_{k,(K)}^i$ by K rounds of exchanging information with neighboring nodes. The ith subvector of the mixed partial gradient $\nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k;t_k)$ associated with node i is given by

$$\nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k;t_k) = \nabla_{t\boldsymbol{y}^i} f^i(\boldsymbol{y}_k^i;t_k) + \nabla_{t\boldsymbol{y}^i} g^{i,i}(\boldsymbol{y}_k^i;t_k) + \sum_{j\in N^i}\nabla_{t\boldsymbol{y}^i} g^{i,j}(\boldsymbol{y}_k^i,\boldsymbol{y}_k^j;t_k). \tag{23}$$

Node i can compute $\nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k;t_k)$ by having access to the decision variables of its neighbors $\boldsymbol{y}_k^j$. In addition, according to the definition of the block diagonal matrix $\mathbf{D}_k$ in (16), its ith

Algorithm 1: Decentralized Prediction at Node i.

Input: The local variable $\boldsymbol{y}_k^i$, the sampling period h, the approximation level K.

1: Compute $\mathbf{D}_k^{ii}$ [cf., (24)]
2: Exchange the variable $\boldsymbol{y}_k^i$ with neighbors $j \in N^i$
3: Compute the local mixed partial gradient $\nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k; t_k)$ [cf., (23)]
4: Compute $\mathbf{B}_k^{ij} := -\nabla_{\boldsymbol{y}^i \boldsymbol{y}^j} g^{i,j}(\boldsymbol{y}_k^i, \boldsymbol{y}_k^j; t_k) \quad$ for $j \in N^i$ [cf., (25)]
5: Compute $\boldsymbol{p}_{k,(0)}^i = -(\mathbf{D}_k^{ii})^{-1} \nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k; t_k)$.
6: **for** $\tau = 0, 1, 2, \ldots, K-1$ **do**
7: Exchange prediction direction $\boldsymbol{p}_{k,(\tau)}^i$ with neighbors $j \in N^i$
8: Compute the recursion [cf., (22)]

$$\boldsymbol{p}_{k,(\tau+1)}^i = -(\mathbf{D}_k^{ii})^{-1} \left(\sum_{j \in N^i} \mathbf{B}_k^{ij} \boldsymbol{p}_{k,(\tau)}^j + \nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k; t_k) \right)$$

9: **end for**
10: Predict the next trajectory $\boldsymbol{y}_{k+1|k}^i = \boldsymbol{y}_k^i + h\, \boldsymbol{p}_{k,(K)}^i$ [cf., (26)]

Output: The predicted variable $\boldsymbol{y}_{k+1|k}^i$.

Algorithm 2: Approximate Prediction at Node i.

Input: The local variable $\boldsymbol{y}_k^i$, the sampling period h, the approximation level K.

1: Compute $\mathbf{D}_k^{ii}$ [cf. (24)]
2: Exchange the variable $\boldsymbol{y}_k^i$ with neighbors $j \in N^i$
3: Compute the approximate local mixed partial gradient $\tilde{\nabla}_{t\boldsymbol{y}} F_k^i$ [cf. (27)]
4: Compute $\mathbf{B}_k^{ij} := -\nabla_{\boldsymbol{y}^i \boldsymbol{y}^j} g^{i,j}(\boldsymbol{y}_k^i, \boldsymbol{y}_k^j; t_k) \quad$ for $j \in N^i$ [cf. (25)]
5: Compute $\tilde{\boldsymbol{p}}_{k,(0)}^i = -(\mathbf{D}_k^{ii})^{-1} \tilde{\nabla}_{t\boldsymbol{y}} F_k^i$
6: **for** $\tau = 0, 1, 2, \ldots, K-1$ **do**
7: Exchange prediction direction $\tilde{\boldsymbol{p}}_{k,(\tau)}^i$ with neighbors $j \in N^i$
8: Compute the recursion [cf. (22)]

$$\tilde{\boldsymbol{p}}_{k,(\tau+1)}^i = -(\mathbf{D}_k^{ii})^{-1} \left(\sum_{j \in N^i} \mathbf{B}_k^{ij} \tilde{\boldsymbol{p}}_{k,(\tau)}^j + \tilde{\nabla}_{t\boldsymbol{y}} F_k^i \right)$$

9: **end for**
10: Predict the next trajectory $\boldsymbol{y}_{k+1|k}^i = \boldsymbol{y}_k^i + h\, \tilde{\boldsymbol{p}}_{k,(K)}^i$ [cf. (26)]

Output: The predicted variable $\boldsymbol{y}_{k+1|k}^i$.

block can be written as

$$\mathbf{D}_k^{ii} := \nabla_{\boldsymbol{y}^i \boldsymbol{y}^i} f^i(\boldsymbol{y}_k^i; t_k) + \nabla_{\boldsymbol{y}^i \boldsymbol{y}^i} g^{i,i}(\boldsymbol{y}_k^i; t_k) + \sum_{j \in N^i} \nabla_{\boldsymbol{y}^i \boldsymbol{y}^i} g^{i,j}(\boldsymbol{y}_k^i, \boldsymbol{y}_k^j; t_k) \tag{24}$$

which is available at node i, after receiving $\boldsymbol{y}_k^j$. These observations imply that the initial prediction direction $\boldsymbol{p}_{k,(0)}^i = (\mathbf{D}_k^{ii})^{-1} \nabla_{t\boldsymbol{y}} F^i(\boldsymbol{y}_k; t_k)$ can be computed locally at node i. Further, the blocks of the neighbor sparse matrix $\mathbf{B}_k$ are given by

$$\mathbf{B}_k^{ij} := -\nabla_{y^i y^j} g^{i,j}(y_k^i, y_k^j; t_k), \quad \text{for } j \in N^i \tag{25}$$

which are available at node i. Therefore, node i can compute the recursion in (22) by having access to the τth level approximate prediction direction $\boldsymbol{p}_{k,(\tau)}^j$ of its neighbors $j \in N^i$. After the K rounds of communication with neighboring nodes, to predict the local variable $\boldsymbol{y}_{k+1|k}^i$ at step t_k, node i executes the local update

$$\boldsymbol{y}_{k+1|k}^i = \boldsymbol{y}_k^i + h\, \boldsymbol{p}_{k,(K)}^i \,. \tag{26}$$

Thus, the prediction step in (20) yields a decentralized protocol summarized in Algorithm 1.

B. Time-Derivative Approximation

In practical settings, knowledge of how the function F changes in time is unavailable. This issue may be mitigated by estimating the term $\nabla_{t\boldsymbol{y}} F(\boldsymbol{y}; t)$ via a first-order backward derivative: Let $\tilde{\nabla}_{t\boldsymbol{y}} F_k$ be an approximate version of $\nabla_{t\boldsymbol{y}} F(\boldsymbol{y}_k; t_k)$ computed as a first-order backward derivative

$$\tilde{\nabla}_{t\boldsymbol{y}} F_k = (\nabla_{\boldsymbol{y}} F(\boldsymbol{y}_k; t_k) - \nabla_{\boldsymbol{y}} F(\boldsymbol{y}_k; t_{k-1}))/h \,. \tag{27}$$

The approximation $\tilde{\nabla}_{t\boldsymbol{y}} F_k$ requires only information of the previous discrete time slot. Using (27), we can approximate the prediction direction as

$$\tilde{\boldsymbol{p}}_{k,(K)} := -\mathbf{H}_{k,(K)}^{-1} \tilde{\nabla}_{t\boldsymbol{y}} F_k \,. \tag{28}$$

This may be obtained in a decentralized way via K rounds of communication among neighboring nodes, which may be established as a trivial extension of Proposition 1. Algorithm 1 may be modified to instead make use of the decentralized approximate prediction step in (28), as done in Algorithm 2. Once we obtain this local prediction of the optimizer at the next time t_{k+1}, using information at the current time t_k, the problem (3) is sampled at time t_{k+1}. We make use of this new information in the correction step, as discussed next.

C. Decentralized Correction Step

The predicted variable $\boldsymbol{y}_{k+1|k}$ [cf., (21)] is then corrected via (10) by making use of the objective at time t_{k+1}. Different correction strategies give rise to different correction updates, whose relative merits depend on the application domain at hand. We present two distinct correction steps next.

1) Gradient Correction Step: After the objective at time t_{k+1} is observed, we may execute the correction step (10) with $\boldsymbol{c}_{k+1} = -\nabla_{\boldsymbol{y}} F(\boldsymbol{y}_{k+1|k}; t_{k+1})$, resulting in

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_{k+1|k} - \gamma \nabla_{\boldsymbol{y}} F(\boldsymbol{y}_{k+1|k}; t_{k+1}) \tag{29}$$

which is a gradient correction step. This step is computable in a decentralized fashion since the local component of the gradient

Algorithm 3. Decentralized Gradient Correction at Node i:

Input: The local predicted variable $\boldsymbol{y}^i_{k+1|k}$. The step size γ.

1: Exchange the predicted variable $\boldsymbol{y}^i_{k+1|k}$ with neighbors $j \in N^i$
2: Observe $F^i(\cdot; t_{k+1})$, find $\boldsymbol{c}^i_{k+1} = -\nabla_{\boldsymbol{y}} F^i(\boldsymbol{y}_{k+1|k}; t_{k+1})$ [cf., (30)]
3: Correct the trajectory $\boldsymbol{y}^i_{k+1} = \boldsymbol{y}^i_{k+1|k} + \gamma \boldsymbol{c}^i_{k+1}$ [cf., (31)]

Output: The corrected variable $\boldsymbol{y}^i_{k+1}$.

$F(\boldsymbol{y}_{k+1|k}; t_{k+1})$ at node i is given by

$$\nabla_{\boldsymbol{y}} F^i(\boldsymbol{y}_{k+1|k}; t_{k+1}) = \nabla_{\boldsymbol{y}^i} f^i(\boldsymbol{y}^i_{k+1|k}; t_{k+1}) + \nabla_{\boldsymbol{y}^i} g^{i,i}(\boldsymbol{y}^i_{k+1|k}; t_{k+1}) + \sum_{j \in N^i} \nabla_{\boldsymbol{y}^i} g^{i,j}(\boldsymbol{y}^i_{k+1|k}, \boldsymbol{y}^j_{k+1|k}; t_{k+1}). \tag{30}$$

To implement the expression in (30), node i only requires access to the decision variables $\boldsymbol{y}^j_{k+1|k}$ of its neighbors $j \in N^i$. Thus, if nodes exchange their predicted variable $\boldsymbol{y}^i_{k+1|k}$ with their neighbors they can compute the local correction direction $\boldsymbol{c}^i_{k+1}$ as in (30) and update their predicted variable $\boldsymbol{y}^i_{k+1|k}$ as

$$\boldsymbol{y}^i_{k+1} = \boldsymbol{y}^i_{k+1|k} + \gamma \boldsymbol{c}^i_{k+1}. \tag{31}$$

We call DPC-G as the decentralized prediction-correction method that uses gradient descent in the correction step (Algorithm 3) and the *exact* prediction step (Algorithm 1) in the prediction step. We call DAPC-G as the decentralized approximate prediction-correction method that uses gradient descent in the correction step (Algorithm 3) and the *approximate* prediction step (Algorithm 2) in the prediction step. Both DPC-G and DAPC-G require $K+2$ communication rounds among neighboring nodes per time step.

2) Newton Correction Step: The correction step in (10) could also be considered as a Newton step if we used $\boldsymbol{c}_{k+1} = -\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_{k+1|k}; t_{k+1})^{-1} \nabla_{\boldsymbol{y}} F(\boldsymbol{y}_{k+1|k}; t_{k+1})$. However, as in the discussion regarding the prediction step, computation of the Hessian inverse $\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_{k+1|k}; t_{k+1})^{-1}$ requires global communication. Consequently, we approximate the Hessian inverse $\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_{k+1|k}; t_{k+1})^{-1}$ by truncating its Taylor series as in (19). To be more precise, we define $\mathbf{H}^{-1}_{k+1|k,(K')}$ as the K'th level approximation of the Hessian inverse as

$$\mathbf{H}^{-1}_{k+1|k,(K')} = \mathbf{D}^{-1/2}_{k+1|k} \sum_{\tau=0}^{K'} \left(\mathbf{D}^{-1/2}_{k+1|k} \mathbf{B}_{k+1|k} \mathbf{D}^{-1/2}_{k+1|k} \right)^{\tau} \mathbf{D}^{-1/2}_{k+1|k} \tag{32}$$

where the matrices $\mathbf{D}_{k+1|k}$ and $\mathbf{B}_{k+1|k}$ are defined as

$$\mathbf{D}_{k+1|k} := \nabla_{\boldsymbol{yy}} f(\boldsymbol{y}_{k+1|k}; t_{k+1}) + \operatorname{diag}[\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_{k+1|k}; t_{k+1})] \tag{33a}$$

$$\mathbf{B}_{k+1|k} := \operatorname{diag}[\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_{k+1|k}; t_{k+1})] - \nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_{k+1|k}; t_{k+1}) . \tag{33b}$$

Algorithm 4: Decentralized Newton Correction at Node i.

Input: The local predicted variable $\boldsymbol{y}^i_{k+1|k}$. The approximation level K'. The step size γ.

1: Exchange the predicted variable $\boldsymbol{y}^i_{k+1|k}$ with neighbors $j \in N^i$
2: Observe $F^i(\cdot; t_{k+1})$, compute $\nabla_{\boldsymbol{y}} F^i(\boldsymbol{y}_{k+1|k}; t_{k+1})$ [cf., (30)]
3: Compute matrices $\mathbf{D}^{ii}_{k+1|k}$ and $\mathbf{B}^{ij}_{k+1|k}$, $j \in N^i$ as

$$\mathbf{D}^{ii}_{k+1|k} := \nabla_{\boldsymbol{y}^i \boldsymbol{y}^i} f^i(\boldsymbol{y}^i_{k+1|k}; t_{k+1}) + \nabla_{\boldsymbol{y}^i \boldsymbol{y}^i} g^{i,i}(\boldsymbol{y}^i_{k+1|k}; t_{k+1}) + \sum_{j \in N^i} \nabla_{\boldsymbol{y}^i \boldsymbol{y}^i} g^{i,j}(\boldsymbol{y}^i_{k+1|k}, \boldsymbol{y}^j_{k+1|k}; t_{k+1})$$

$$\mathbf{B}^{ij}_{k+1|k} := -\nabla_{\boldsymbol{y}^i \boldsymbol{y}^j} g^{i,j}(\boldsymbol{y}^i_{k+1|k}, \boldsymbol{y}^j_{k+1|k}; t_{k+1})$$

4: Compute $\boldsymbol{c}^i_{k+1,(0)} = -(\mathbf{D}^{ii}_{k+1|k})^{-1} \nabla_{\boldsymbol{y}} F^i(\boldsymbol{y}_{k+1|k}; t_{k+1})$
5: **for** $\tau = 0, 1, 2, \ldots, K'-1$ **do**
6: Exchange correction step $\boldsymbol{c}^i_{k,(\tau)}$ with neighboring nodes $j \in N^i$
7: Compute $\boldsymbol{c}^i_{k+1,(\tau+1)}$ as

$$\boldsymbol{c}^i_{k+1,(\tau+1)} = -(\mathbf{D}^{ii}_{k+1|k})^{-1} \left(\sum_{j \in N^i} \mathbf{B}^{ij}_{k+1|k} \boldsymbol{c}^j_{k+1,(\tau)} + \nabla_{\boldsymbol{y}} F^i(\boldsymbol{y}_{k+1|k}; t_{k+1}) \right)$$

8: **end for**
9: Correct the trajectory prediction $\boldsymbol{y}^i_{k+1} = \boldsymbol{y}^i_{k+1|k} + \gamma \boldsymbol{c}^i_{k+1,(K')}$

Output: The corrected variable $\boldsymbol{y}^i_{k+1}$.

Notice that the only difference between the decomposition matrices $\mathbf{D}_{k+1|k}$ and $\mathbf{B}_{k+1|k}$ for the correction step and the matrices $\mathbf{D}_k$ and $\mathbf{B}_k$ for the prediction step is the arguments for the inputs $\boldsymbol{y}$ and t. The prediction matrices $\mathbf{D}_k$ and $\mathbf{B}_k$ are evaluated for the function $F(.; t_k)$ and the variable $\boldsymbol{y}_k$, while the correction matrices are evaluated for the function $F(.; t_{k+1})$ and the variable $\boldsymbol{y}_{k+1|k}$.

Thus, we can approximate the exact Hessian inverse $\nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_{k+1|k}; t_{k+1})^{-1}$ with $\mathbf{H}^{-1}_{k+1|k,(K')}$ as in (32) and apply the correction step as

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_{k+1|k} - \gamma \, \mathbf{H}^{-1}_{k+1|k,(K')} \nabla_{\boldsymbol{y}} F(\boldsymbol{y}_{k+1|k}; t_{k+1}) \tag{34}$$

which requires K' exchanges of information among neighboring nodes. In practice, one can use the same algorithm for the prediction direction $\boldsymbol{p}_{k,(K)}$ to compute the correction direction $\boldsymbol{c}_{k,(K')} := -\mathbf{H}^{-1}_{k+1|k,(K')} \nabla_{\boldsymbol{y}} F(\boldsymbol{y}_{k+1|k}; t_{k+1})$, where now the gradient takes the place of the time derivative.

We call DPC-N as the decentralized prediction-correction method that uses Newton descent in the correction step (Algorithm 4) and the *exact* prediction step

TABLE I
COMMUNICATION REQUIREMENTS FOR THE PRESENTED ALGORITHMS

Method		Comms.	Vars.	Vars. Communicated
DPC-G/ DAPC-G	Pred.	$K+1$	$\boldsymbol{p}_k^i, \boldsymbol{y}_k^i$	$(K+1)p$
	Corr.	1	$\boldsymbol{y}_{k+1\|k}^i$	p
DPC-N/ DAPC-N	Pred.	$K+1$	$\boldsymbol{p}_k^i, \boldsymbol{y}_k^i$	$(K+1)p$
	Corr.	$K'+1$	$\boldsymbol{c}_k^i, \boldsymbol{y}_{k+1\|k}^i$	$(K'+1)p$

(Algorithm 1) in the prediction step. We call DAPC-N as the decentralized approximate prediction-correction method that uses Newton descent in the correction step (Algorithm 4) and the *approximate* prediction step (Algorithm 2) in the prediction step. Both DPC-N and DAPC-N require $K+K'+2$ rounds of communication per iteration.

For the reader's ease, we report in Table I the total communication counts per iteration for the presented algorithms. In particular, we report the amount of communication rounds required among the neighboring nodes, as well as the variables that have to be transmitted and the total number of scalar variables to be sent (per neighbor).

IV. CONVERGENCE ANALYSIS

We continue by establishing the convergence of the methods presented in Section III. In particular, we show that as time passes the sequence $\{\boldsymbol{y}_k\}$ approaches a neighborhood of the optimal trajectory $\boldsymbol{y}^*(t_k)$ at discrete time instances t_k. To establish our results, we require the following conditions.

Assumption 1: The local functions f^i are twice differentiable and the eigenvalues of their Hessians $\nabla_{\boldsymbol{y}^i\boldsymbol{y}^i} f^i(\boldsymbol{y}^i;t)$ for all i are contained in a compact interval $[m,M]$ with $m>0$. Hence, the aggregate function $f(\boldsymbol{y};t):=\sum_{i\in V} f^i(\boldsymbol{y}^i;t)$ has a uniformly bounded spectrum, i.e.

$$m\boldsymbol{I} \preceq \nabla_{\boldsymbol{y}\boldsymbol{y}} f(\boldsymbol{y};t) \preceq M\boldsymbol{I}. \tag{35}$$

Assumption 2: The functions $g^{i,i}(\boldsymbol{y}^i;t)$ and $g^{i,j}(\boldsymbol{y}^i,\boldsymbol{y}^j;t)$ are twice differentiable. The Hessian of the aggregate in (2), denoted as $\nabla_{\boldsymbol{y}\boldsymbol{y}} g(\boldsymbol{y};t)$, is block diagonally dominant [35], i.e., for all i

$$\left\|\nabla_{\boldsymbol{y}^i\boldsymbol{y}^i} g(\boldsymbol{y}^i,\boldsymbol{y}^j;t)^{-1}\right\|^{-1} \geqslant \sum_{j=1,j\neq i}^{n} \left\|\nabla_{\boldsymbol{y}^i\boldsymbol{y}^j} g^{i,j}(\boldsymbol{y}^i,\boldsymbol{y}^j;t)\right\| \tag{36}$$

where by definition $\nabla_{\boldsymbol{y}^i\boldsymbol{y}^i} g(\boldsymbol{y}^i,\boldsymbol{y}^j;t)=\nabla_{\boldsymbol{y}^i\boldsymbol{y}^i} g^{i,i}(\boldsymbol{y}^i;t)+\nabla_{\boldsymbol{y}^i\boldsymbol{y}^i} g^{i,j}(\boldsymbol{y}^i,\boldsymbol{y}^j;t)$. The block diagonal element $\nabla_{\boldsymbol{y}^i\boldsymbol{y}^i} g(\boldsymbol{y}^i,\boldsymbol{y}^j;t)$ has eigenvalues contained in a compact interval $[\ell/2, L/2]$ with $\ell>0$.

Assumption 3: The derivatives of the global cost $F(\boldsymbol{y};t)$ defined in (3) are bounded for all $\boldsymbol{y}\in\mathbb{R}^{np}$ and $t\geqslant 0$ as

$$\begin{aligned} &\|\nabla_{t\boldsymbol{y}} F(\boldsymbol{y};t)\|\leqslant C_0,\ \|\nabla_{\boldsymbol{y}\boldsymbol{y}\boldsymbol{y}} F(\boldsymbol{y};t)\|\leqslant C_1\\ &\|\nabla_{\boldsymbol{y}t\boldsymbol{y}} F(\boldsymbol{y};t)\|\leqslant C_2,\ \|\nabla_{tt\boldsymbol{y}} F(\boldsymbol{y};t)\|\leqslant C_3\,. \end{aligned} \tag{37}$$

From the bounds on the eigenvalues of Hessians $\nabla_{\boldsymbol{y}\boldsymbol{y}} f(\boldsymbol{y};t)$ and $\nabla_{\boldsymbol{y}\boldsymbol{y}} g(\boldsymbol{y};t)$ in Assumptions 1 and 2, respectively, and from the block diagonal Gerschgorin Circle Theorem [35] it follows that the spectrum of $\nabla_{\boldsymbol{y}\boldsymbol{y}} g(\boldsymbol{y};t)$ lies in the compact set $[0,L]$, and the one of the Hessian of the global cost $\nabla_{\boldsymbol{y}\boldsymbol{y}} F(\boldsymbol{y};t)$ uniformly satisfies

$$m\,\mathbf{I} \preceq \nabla_{\boldsymbol{y}\boldsymbol{y}} F(\boldsymbol{y};t) \preceq (L+M)\,\mathbf{I}\,. \tag{38}$$

Assumptions 1 and 2, besides guaranteeing that the problem stated in (3) is strongly convex and has a *unique* solution for each time instance, imply that the Hessian $\nabla_{\boldsymbol{y}\boldsymbol{y}} F(\boldsymbol{y};t)$ is invertible. Moreover, the higher order derivative bounds imply the Lipschitz continuity of the gradients, Hessians, and mixed partial derivatives of the Hessians. These conditions, in addition to higher order derivative conditions on F, as in Assumption 3, frequently appear in the analysis of methods for time-varying optimization, and are required to establish convergence [13], [18], [19].

Assumptions 1 and 3 are sufficient to show that the solution *mapping* $t\mapsto \boldsymbol{y}^*(t)$ is single-valued and locally Lipschitz continuous in t, and in particular

$$\|\boldsymbol{y}^*(t_{k+1})-\boldsymbol{y}^*(t_k)\|\leqslant \frac{1}{m}\|\nabla_{t\boldsymbol{y}} F(\boldsymbol{y};t)\|(t_{k+1}-t_k)\leqslant \frac{C_0 h}{m} \tag{39}$$

see, for example, [36, Th. 2F.10]. This gives us a link between the sampling period h and the allowed variations in the optimizers. This also gives a better understanding on the time-varying assumptions on the uniform boundedness of the time derivatives of the gradient $\nabla_{t\boldsymbol{y}} F(\boldsymbol{y};t)$ and $\nabla_{tt\boldsymbol{y}} F(\boldsymbol{y};t)$. In particular the bounds C_0 and C_3 require that the change and the rate of change of the optimizer be bounded. If the optimizer were the position of a moving target to be estimated, then C_0 and C_3 would be a bound on its velocity and acceleration. Finally, the bound on $\nabla_{\boldsymbol{y}t\boldsymbol{y}} F(\boldsymbol{y};t)$ means that the quantity $\nabla_{t\boldsymbol{y}} F(\boldsymbol{y};t)$ is Lipschitz continuous, w.r.t., $\boldsymbol{y}$ uniformly in t. That is to say that close by points $\boldsymbol{y}$ and $\boldsymbol{y}'$ need to have similar gradient time derivatives: e.g., if the target position is perturbed by a small amount $\delta\boldsymbol{y}$, then its velocity is perturbed by an amount not bigger than $C_2\delta\boldsymbol{y}$.

A. Discrete Sampling Error

We start the convergence analysis by deriving an upper bound on the norm of the approximation error $\boldsymbol{\Delta}_k\in\mathbb{R}^{np}$ that we estimate through a Taylor approximation in (13). The error is defined as the difference between the predicted $\boldsymbol{y}_{k+1|k}$ in (14) (with $\boldsymbol{y}_k=\boldsymbol{y}^*(t_k)$) and the exact prediction $\boldsymbol{y}^*(t_{k+1})$, starting from the same initial condition $\boldsymbol{y}^*(t_k)$, i.e.

$$\boldsymbol{\Delta}_k := \boldsymbol{y}_{k+1|k}-\boldsymbol{y}^*(t_{k+1}). \tag{40}$$

In the following proposition, we upper bound the norm $\|\boldsymbol{\Delta}_k\|$ of the discretization error, which encodes the error due to the prediction step and is central to all our convergence results.

Proposition 2: Let Assumptions 1–3 hold true. Define the discretization constant Δ as $\Delta=(C_0^2C_1)/2m^3+(C_0C_2)/m^2+(C_3)/2m$. The norm of $\boldsymbol{\Delta}_k$ in (40) is upper bounded by

$$\|\boldsymbol{\Delta}_k\|\leqslant \Delta\, h^2 = O(h^2). \tag{41}$$

Proposition 2, which is established as [10, Proposition 1], states that the norm of the discrete sampling error $\|\boldsymbol{\Delta}_k\|$ is bounded above by a constant, which is in the order of $O(h^2)$.

A second source of error to take into account, when studying the asymptotic behavior of the algorithms in Section III, is the error due to approximating the Hessian inverse by a truncated Taylor expansion in (19). We bound this error as a function of the approximation level K, which is the number of communication rounds among neighboring nodes.

Proposition 3: Under Assumptions 1 and 2, the Kth order approximate inverse Hessian in (19) is well-defined. In addition, its eigenvalues are upper bounded as

$$\|\mathbf{H}_{k,(K)}^{-1}\| \leqslant H := \frac{m + L/2}{m(m + \ell/2)}. \tag{42}$$

Furthermore, if we define the error of the Hessian inverse approximation as $e_k = \|\mathbf{I} - \nabla_{\boldsymbol{yy}} F(\boldsymbol{y}_k; t_k)\mathbf{H}_{k,(K)}^{-1}\|$, the error e_k is bounded above as

$$e_k \leqslant \varrho^{K+1}, \quad \text{where } \varrho := (L/2)/(m + L/2). \tag{43}$$

Proof: See Appendix A. ■

Besides quantifying the error coming from approximating the Hessian inverse, Proposition 3 provides tradeoffs between communication cost and convergence accuracy. It shows that a larger K leads to more accurate approximation of the Hessian inverse at the price of more communications.

B. Gradient Tracking Convergence

In the following theorem, we establish that the sequence generated by the DPC-G and DAPC-G algorithms asymptotically converges to a neighborhood of the optimal trajectory whose radius depends on the discretization error.

Theorem 1: Consider the sequence $\{\boldsymbol{y}_k\}$ generated by the DPC-G or DAPC-G algorithm, which uses Algorithm 1 (or 2) as prediction step and Algorithm 3 as correction step. Let Assumptions 1–3 hold and define constants ρ and σ as

$$\begin{aligned} \rho &:= \max\{|1 - \gamma m|, |1 - \gamma(L + M)|\}, \\ \sigma &:= 1 + h\left[\frac{C_0 C_1}{m^2} + \frac{C_2}{m}\right]. \end{aligned} \tag{44}$$

Further, recall the definition of ϱ in (43) and define the function $\Gamma : (0, 1) \times \mathbb{N} \to \mathbb{R}$ as $\Gamma(\varrho, K) = (C_0/m)\varrho^{K+1}$. Choose the step size as

$$\gamma < 2/(L + M) \tag{45}$$

so that $\rho < 1$. Then,

1) for any sampling period h, the sequence $\{\boldsymbol{y}_k\}$ converges to $\boldsymbol{y}^*(t_k)$ Q-linearly up to a bounded error, as

$$\limsup_{k\to\infty} \|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| = O(h) + O(h\Gamma(\varrho, K)) + O(h^2) \tag{46}$$

2) if the sampling period h is chosen such that

$$h < \left[\frac{C_0 C_1}{m^2} + \frac{C_2}{m}\right]^{-1} (\rho^{-1} - 1) \tag{47}$$

then the sequence $\{\boldsymbol{y}_k\}$ converges to $\boldsymbol{y}^*(t_k)$ Q-linearly up to a bounded error as

$$\limsup_{k\to\infty} \|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| = O(h\Gamma(\varrho, K)) + O(h^2). \tag{48}$$

Proof: See Appendix B, where the error bounds and convergence rate constant are explicitly computed in terms of the functional bounds of Assumptions 1–3. ■

Theorem 1 establishes the convergence properties of DPC-G and DAPC-G for particular parameter choices. In both cases, the linear convergence to a neighborhood is shown, provided the step size satisfies $\gamma < 2/(L + M)$. Moreover, the accuracy of convergence depends on the choice of the sampling period h, and for any sampling period, the result in (46) holds. In this case, the accuracy of convergence is of the order $O(h)$. If the sampling period h is chosen such that $\rho\sigma < 1$ (that is (47) holds), then the result in (48) is valid.

If $\rho\sigma < 1$ is satisfied and the approximation level K is chosen sufficiently large, then $\Gamma(\varrho, K)$ is negligible and we regain an error bound of $O(h^2)$, which is compatible with centralized algorithms [10].

C. Newton Tracking Convergence

We turn to analyzing the DPC-N and DAPC-N algorithms.

Theorem 2: Denote $\{\boldsymbol{y}_k\}$ as the sequence generated by the DPC-N or DAPC-N method, which, respectively, uses Algorithm 1 or 2 as its prediction, and Algorithm 4 as its correction. Let Assumptions 1–3 hold and fix K and K' as the Hessian inverse approximation levels for the prediction and correction steps, respectively, with the function Γ defined as in Theorem 1. Fix the step size as $\gamma \in (0, 1]$. There exist bounds $\bar{K}$, $\bar{h}$, and $\bar{R}$, such that if the sampling rate h is chosen as $h \leqslant \bar{h}$, K and K' are chosen as $K, K' \geqslant \bar{K}$, and the initial optimality gap satisfies $\|\boldsymbol{y}_0 - \boldsymbol{y}^*(t_0)\| \leqslant \bar{R}$, then $\{\boldsymbol{y}_k\}$ converges Q-linearly to the solution trajectory $\boldsymbol{y}^*(t_k)$ up to a bounded error as

$$\begin{aligned} \limsup_{k\to\infty} \|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| &= O(h\Gamma(\varrho, K)[\gamma\Gamma(\varrho, K') + 1 - \gamma]) \\ &\quad + O(h^2[\gamma\Gamma(\varrho, K') + \gamma\Gamma(\varrho, K)^2 + 1 - \gamma]) \\ &\quad + O(h^3\gamma\Gamma(\varrho, K)) + O(h^4\gamma). \end{aligned} \tag{49}$$

In addition, if the step size γ is chosen arbitrarily small, the attraction region $\bar{R}$ can be made arbitrarily large.

Proof: See Appendix C. The proof is constructive, thus we also characterize the bounds on the sampling period, approximation levels, the attraction region, and finally, the constants in the asymptotic error and in the linear convergence rate. ■

Theorem 2 states that DPC-N/DAPC-N converge to a bounded tracking error defined in (49) once the algorithm reaches an attraction region. The error bound in (49) depends, as expected, on the sampling period h and the approximation levels K and K'. In the worst case, the asymptotic error floor will be of the order $O(h)$. However, in some cases, we may achieve tighter tracking guarantees. For example, if the approximation level K and K' are chosen sufficiently large,

then the terms $\Gamma(\rho, K)$ and $\Gamma(\rho, K')$ are negligible, yielding

$$\limsup_{k\to\infty} \|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| = O(h^2[1-\gamma]) + O(h^4\,\gamma)\,. \quad (50)$$

This is to say that the asymptotic dependence of the error of DPC-N/DAPC-N on the sampling period h varies from a worst-case $O(h)$ to as tight as $O(h^4)$ (for the selection $\gamma = 1$).

Remark 1: (Step-size choice) The step-size choice of the presented Newton correction methods affects the convergence attraction region and the convergence speed. For large enough K, K' and small enough h, if we choose $\gamma = 1$, we obtain a standard Newton method with convergence region $\bar{R} = 2m/C_1\sigma^2$, the fastest convergence speed, and smallest asymptotical error $O(h^4)$. This convergence region is larger depending on how small C_1 is: for quadratic functions $C_1 = 0$, and the convergence is global. If we choose $\gamma \ll 1$, then the convergence region is $\bar{R} = 2m(\tau - 1 + \gamma)/\gamma C_1\sigma^2$, where τ, with $1-\gamma < \tau < 1$, is the linear convergence rate [Cf., Appendix C]. This means that the attraction region can be made arbitrarily big, while the convergence rate is made smaller and smaller, and the asymptotical error is $O(h^2)$. Finally, an interesting choice is $\gamma = h \leqslant 1$: when h is sufficiently small, then the Newton prediction/correction approximate well a continuous-time algorithm and the convergence (albeit made slow) is global. In practice, the choice of the step size depends on the application at hand. One can even decide to run the DPC-G algorithm till convergence and then switch to DPC-N as an hybrid scheme, or to adopt an increasing step size selection. These extensions are left as future research.

V. Numerical Evaluation

We turn to studying the empirical validity of the performance guarantees established in Section IV. In particular, we consider the resource allocation problem in a network of interconnected devices, as in Example 2 of Section II. As presented in (5), the local objective functions $f^i(\boldsymbol{y}^i; t)$ represent a time-varying utility indicating the quality of transmission at a particular device i and the constraints represent channel rate and capacity constraints. These constraints depend on the connectivity of the network, which is encoded in the augmented incidence matrix $\boldsymbol{A}$, defined following (5).

By adopting an approximate augmented Lagrangian method, we obtain (6), which is an instance of (3). Consider the case where decisions are the variables $\boldsymbol{y}^i \in \mathbb{R}^{\boldsymbol{p}}$, $p = 10$, for which each local utility $f^i(\boldsymbol{y}^i; t)$ associated with sensor i is given as

$$f^i(\boldsymbol{y}^i; t) = \frac{1}{2}(\boldsymbol{y}^i - \boldsymbol{c}^i(t))^{\mathsf{T}}\mathbf{Q}^i(\boldsymbol{y}^i - \boldsymbol{c}^i(t)) + \sum_{l=1}^{p} \log\left[1 + \exp\left(b^{i,l}(y^{i,l} - d^{i,l}(t))\right)\right] \quad (51)$$

where $\boldsymbol{y}^{i,l}$ indicates the lth component of the ith decision variable $\boldsymbol{y}^i$, while $\mathbf{Q}^i \in \mathbb{R}^{p\times p}, b^{i,l} \in \mathbb{R}, \boldsymbol{c}^i(t) \in \mathbb{R}^{\boldsymbol{p}}, d^{i,l}(t) \in \mathbb{R}$ are (time-varying) parameters. Straightforward computations reveal that the second-order derivative of f^i with respect to $\boldsymbol{y}^i$ is contained in the bounded interval $[\lambda_{\min}(\mathbf{Q}^i), \lambda_{\max}(\mathbf{Q}^i) + \max_l\{(b^{i,l})^2/4\}]$.

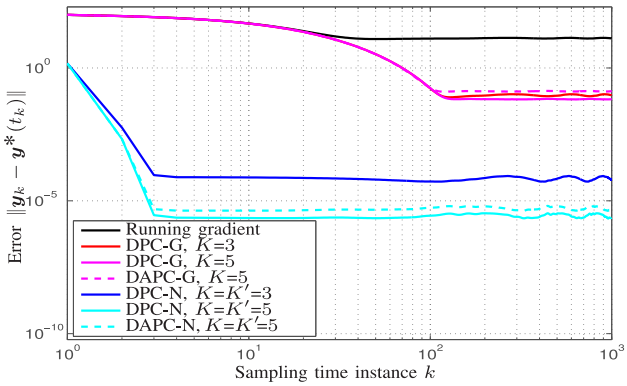

Fig. 1. Error with respect to the sampling time instance k for different algorithms applied to the continuous-time sensor network resource allocation problem (6), with sampling interval $h = 0.1$.

Experimentally, we consider cases were each $\mathbf{Q}^i$ and $b^{i,l}$ are selected uniformly at random, and in particular, $\mathbf{Q}_i = \text{diag}(\mathcal{U}^p_{[1,2]}) + \boldsymbol{v}^i\boldsymbol{v}^{i\,\mathsf{T}}$, with $\boldsymbol{v}_i \sim \mathcal{N}^p_{0,1}$ (that is $\boldsymbol{v}_i$ is a random vector drawn from a Gaussian distribution of mean zero and standard deviation one). With this choice $\mathbf{Q}^i$ is positive definite. In addition $b^{i,j} \sim \mathcal{U}^1_{[-2,2]}$. Finally, $\boldsymbol{c}^i(t)$ and $d^{i,l}(t)$ are the time-varying functions

$$c^{i,l}(t) = 10\cos(\theta_c^{i,l} + \omega\, t), \quad \theta_c^{i,l} \sim \mathcal{U}[0, 2\pi) \quad (52a)$$

$$d^{i,l}(t) = 10\cos(\theta_d^{i,l} + \omega\, t), \quad \theta_d^{i,l} \sim \mathcal{U}[0, 2\pi) \quad (52b)$$

with $\omega = 0.1$. The sensors in the $n = 50$ node wireless network are deployed randomly in the area $[-1, 1]^2$ and can communicate if they are closer than a range of $r = 2.5\sqrt{2}/\sqrt{n}$, which generates a network of l links. We set the vector of rate and capacity constraints to $\boldsymbol{b} = \mathbf{0}$ yielding a dynamic network flow problem, with approximation level $\beta = \sqrt{20}$.

A. Comparisons in Absolute Terms

We first analyze the behavior of DPC-G, DAPC-G, DPC-N, and DAPC-N with respect to the decentralized running gradient method of [20]. Unless otherwise stated the DPC-N and DAPC-N algorithms run with unitary step size ($\gamma = 1$).

In Fig. 1, we depict how the different algorithms reach convergence as time passes for a fixed sampling interval of $h = 0.1$. Observe that the running gradient method achieves the worst tracking performance of around $\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| \approx 10$, whereas DPC-G for various levels of communication rounds in the prediction step K achieves an error near 10^{-1}. Using second-order information in the correction step, as with DPC-N and DAPC-N, achieves superior performance, with tracking errors of at least $\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| \approx 10^{-5}$. Moreover, the time approximation in DAPC-G and DAPC-N does not degrade significantly the asymptotic error, while the number of communication rounds K and K' play a more dominant role, especially in the case of DPC-N.

We also observe this trend in Fig. 2, where we analyze the behavior varying the sampling period h. We approximate the asymptotic error bound as $\max_{k>\bar{k}}\{\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\|\}$, for a given $\bar{k}$, where we set $\bar{k} = 800$ for $h \geqslant 1/16$ or $\bar{k} = 2000$ for

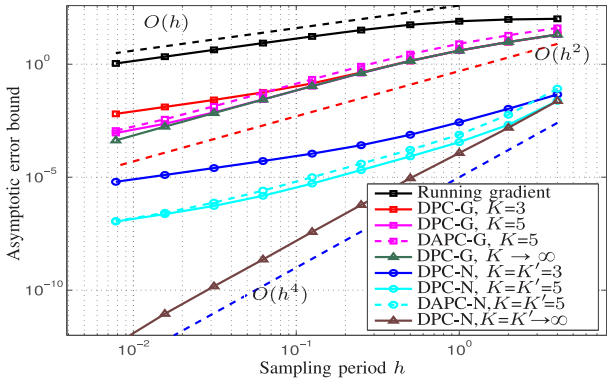

Fig. 2. Asymptotic error bound $\max_{k>\bar{k}}\{\|\boldsymbol{y}_k-\boldsymbol{y}^*(t_k)\|\}$ as compared with the sampling interval h. Dotted straight lines represent error bounds $O(h^r)$ for $r=1,2,4$.

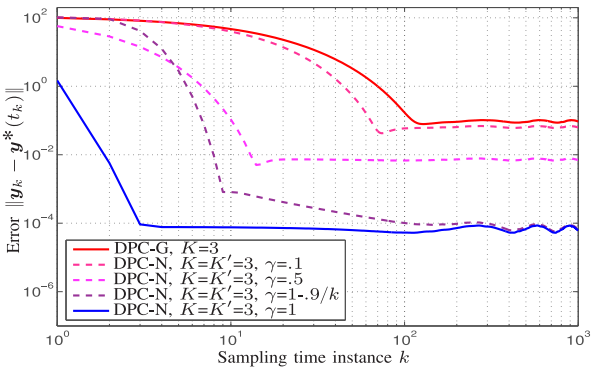

Fig. 3. Error with respect to the sampling time instance k for DPC-G and DPC-N with different step size γ, applied to the continuous-time sensor network resource allocation problem (6), with sampling interval $h=0.1$.

$h<1/16$. We may observe empirical confirmation of the error bounds established by Theorems 1 and 2 in Section IV. In particular, the running gradient has an asymptotic error approximately as $O(h)$, whereas that of DPC-G varies between $O(h)$ and $O(h^2)$ depending on the approximation level K and h. Moreover, DPC-N achieves an asymptotic tracking error varying between $O(h)$ and $O(h^4)$.

In Fig. 3, we depict the behavior in time for different choices of step size γ for DPC-N. As we notice, varying from a small step size $\gamma=0.1=h$ to the biggest one of $\gamma=1$, the convergence becomes faster (yet theoretically more local). An increasing choice of step size as $\gamma=1-.9/k$ seems to combine both larger convergence region, reasonably fast convergence, and small asymptotical error.

Since DPC-N is a computationally more demanding method in terms of communication requirements and computational latency, we study the effect of fixing the former parameter.

B. Comparisons With Fixed Communication Effort

In practice, the communication and computation requirements for each of the nodes of the network will be fixed by hardware and bandwidth constraints. Let us fix the time, as a percentage of the sampling period h, for the prediction and correction step. Let us say that we have at most a time of rh ($r\leqslant .5$) to do prediction and rh to do correction.

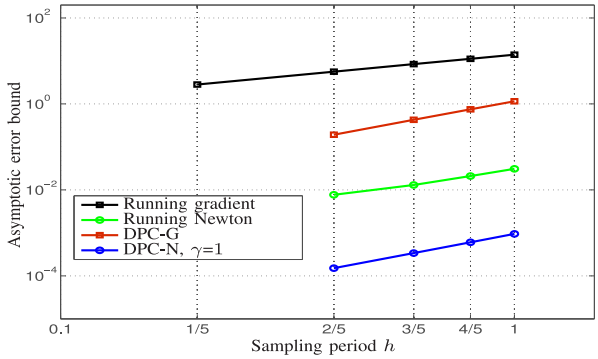

Fig. 4. Asymptotic error with respect to the sampling period h for different algorithms when the number of communication rounds is chosen according to bandwidth constraints as in (53).

Each time a new function is sampled, each of the proposed algorithms will perform a number of correction steps $n_{\mathrm{C}}\geqslant 1$. Each of them will consist of either n_{C} gradient steps, involving each broadcasting p scalar values to the neighbors and receiving pN_i scalar values from them, or n_{C} approximate Newton steps, involving each broadcasting $p(K'+1)$ scalar values to the neighbors and receiving $pN_i(K'+1)$ scalar values.

Once the corrected variable is derived, it can be implemented (e.g., generating the control action). In the remaining time, while waiting for another sampled cost function, the proposed algorithms can perform a prediction step, involving for each node broadcasting $p(K+1)$ scalar values to the neighbors and receiving $pN_i(K+1)$ scalar values from them. For the running schemes, there is no prediction, but we assume here that the variables are further optimized by other extra correction steps, and hence, start at the next time with a better initialization. These further correction steps, say n_{EC}, can be gradient or Newton.

Define $\bar{t}$ to be the time required for one round of broadcasting and receiving data from and to the neighbors and assume that it is the same for each node and it scales linearly with the number of communication rounds K and K' (since it has to be done sequentially). Suppose, as empirically observed, that the computation time for the nodes is negligible, w.r.t., the communication time. In this context, the number of correction and prediction rounds can be chosen according to the constraints on time

$$\text{(RG)}\qquad n_{\mathrm{C}}\bar{t}=rh \qquad n_{\mathrm{EC}}\bar{t}=rh \tag{53a}$$

$$\text{(RN)}\qquad n_{\mathrm{C}}(K'+1)\bar{t}=rh \qquad n_{\mathrm{EC}}(K+1)\bar{t}=rh \tag{53b}$$

$$\text{(DPC-G)}\qquad n_{\mathrm{C}}\bar{t}=rh \qquad (K+1)\bar{t}=rh \tag{53c}$$

$$\text{(DPC-N)}\qquad n_{\mathrm{C}}(K'+1)\bar{t}=rh \qquad (K+1)\bar{t}=rh \tag{53d}$$

where RG indicates the running gradient method and RN the running Newton. In the following simulation, we fix $r=0.5$, $K=K'$ and n_{C}, n_{EC} to be 1 for RN and DPC-N. We fix $\bar{t}=1/10$ s (for bigger values of p, that is the dimension of the decision variable, this time will be longer, and vice versa).

In Fig. 4, we report the results in terms of asymptotical error when optimizing the number of communication rounds

TABLE II
SUMMARY OF PROPOSED METHODS AND CONVERGENCE RESULTS

Method	DPC-G	DAPC-G	DPC-N	DAPC-N
Prediction	Alg. 1	Alg. 2	Alg. 1	Alg. 2
Correction	Alg. 3	Alg. 3	Alg. 4	Alg. 4
Best error bound	$O(h^2)$	$O(h^2)$	$O(h^4)$	$O(h^4)$

according to (53), for different sampling periods. In this context, e.g., for $h = 1$ s, we can run RG with $n_{\mathrm{C}} = n_{\mathrm{EC}} = 5$ correction and extra correction rounds, RN with $K = K' = 4$ communication rounds for correction and extra correction, DPC-G with $K = 4$ communication rounds for prediction and $n_{\mathrm{C}} = 5$ rounds of correction, and DPC-N with $K = K' = 4$ communication rounds for prediction and correction, respectively. For the other values of sampling period h, similar calculations give us the optimized values for n_{C}, n_{EC}, K, and K'. As we observe, when the sampling period is big enough so that DPC-N is implementable, then it seems to be the best strategy to go for.[2] We also notice that running gradient, even with the extra correction steps, should be avoided unless all the remaining algorithms are unviable (as for $h = 1/5$ s).

VI. CONCLUSION

We considered continuously varying convex programs whose objectives may be decomposed into two parts: A sum of locally available functions at the nodes and a part that is shared between neighboring nodes. To solve this problem and track the solution trajectory, we proposed a decentralized iterative procedure that samples the problem at discrete times. Each node predicts where the solution trajectory will be at the next time via an approximation procedure in which it communicates with its neighbors, and then, corrects this prediction by incorporating information about how the local objective is varying, again via a decentralized local approximation. We developed an extension of this tool which allows for the case when the dynamical behavior of the objective must be estimated.

We established that this decentralized approximate second-order procedure converges to an asymptotic error bound which depends on the length of the sampling interval and the amount of communications in the network. Moreover, we established that this convergence result also applies to the case where time derivatives must be approximated. A summary of the proposed methods and their performance guarantees is given in Table II. Finally, we applied the developed tools to a resource allocation problem in a wireless network, demonstrating its practical utility and its ability to outperform existing running methods by orders of magnitude.

APPENDIX A
PROOF OF PROPOSITION 3

We generalize the proofs of Propositions 2 and 3 in [24] to establish the result. Start by defining the matrix $\hat{\mathbf{D}}_k = \mathrm{diag}[\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_k; t_k)]$, which is positive definite due to Assumption 1. Thus, we can write

$$\begin{aligned}\|\mathbf{D}_k^{-1/2}\mathbf{B}_k\mathbf{D}_k^{-1/2}\| &= \|\mathbf{D}_k^{-1/2}\hat{\mathbf{D}}_k^{1/2}\hat{\mathbf{D}}_k^{-1/2} \\ &\quad \times \mathbf{B}_k\hat{\mathbf{D}}_k^{-1/2}\hat{\mathbf{D}}_k^{1/2}\mathbf{D}_k^{-1/2}\| \\ &\leqslant \|\mathbf{D}_k^{-1/2}\hat{\mathbf{D}}_k^{1/2}\|^2\|\hat{\mathbf{D}}_k^{-1/2}\mathbf{B}_k\hat{\mathbf{D}}_k^{-1/2}\| \quad (54)\end{aligned}$$

where the inequality in (54) is implied by the Cauchy-Schwartz inequality. We proceed to bound both terms on the right-hand side (RHS) of (54), starting with the rightmost term. The matrix $\hat{\mathbf{D}}_k^{-1/2}\mathbf{B}_k\hat{\mathbf{D}}_k^{-1/2}$ is conjugate to the matrix $\hat{\mathbf{D}}_k^{-1}\mathbf{B}_k$, which means that the latter has the same eigenvalues of the former. By construction, the matrix $\mathbf{B}_k\hat{\mathbf{D}}_k^{-1}$ is equivalent to

$$\hat{\mathbf{D}}_k^{-1}\mathbf{B}_k = \mathbf{I} - \hat{\mathbf{D}}_k^{-1}\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_k; t_k). \quad (55)$$

The matrix $\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_k; t_k)$ is block diagonally dominant (Assumption 2); by the definition of the matrix $\hat{\mathbf{D}}_k$, this means

$$\|[\hat{\mathbf{D}}_k^{-1}]^{ii}\|^{-1} \geqslant \sum_{j=1, j\neq i}^{n} \left\|\nabla_{\boldsymbol{y}^i\boldsymbol{y}^j} g^{i,j}(\boldsymbol{y}_k^i, \boldsymbol{y}_k^j; t)\right\| \text{ for all } i. \quad (56)$$

Now consider the matrix $\hat{\mathbf{D}}_k^{-1}\nabla_{\boldsymbol{yy}} g(\boldsymbol{y}_k; t_k)$, by the block Gershgorin Circle theorem [35] its eigenvalues are contained in the circles defined by all the μ's that verify

$$\|(\mathbf{I} - \mu\mathbf{I})^{-1}\|^{-1} \leqslant \sum_{j=1, j\neq i}^{n} \left\|[\hat{\mathbf{D}}_k^{-1}]^{ii}\nabla_{\boldsymbol{y}^i\boldsymbol{y}^j} g^{i,j}(\boldsymbol{y}_k^i, \boldsymbol{y}_k^j; t)\right\| \leqslant 1 \quad (57)$$

where the last inequality comes from (56). Therefore the eigenvalues are contained in the compact set $[0, 2]$. This means that the matrix $\hat{\mathbf{D}}_k^{-1}\mathbf{B}_k$ in (55) has eigenvalues contained in the compact set $[-1, 1]$. Taken with the fact that the Frobenius norm of the matrix $\hat{\mathbf{D}}_k^{-1/2}\mathbf{B}_k\hat{\mathbf{D}}_k^{-1/2}$ is bounded above by its maximum eigenvalue, we have

$$\|\hat{\mathbf{D}}_k^{-1/2}\mathbf{B}_k\hat{\mathbf{D}}_k^{-1/2}\| \leqslant 1. \quad (58)$$

With this bound in place, we shift focus to the first term on the RHS of (54). Note that the matrices $\mathbf{D}_k$ and $\hat{\mathbf{D}}_k$ are both symmetric and positive definite and therefore we can write $\|\mathbf{D}_k^{-1/2}\hat{\mathbf{D}}_k^{1/2}\|^2 = \|\mathbf{D}_k^{-1/2}\hat{\mathbf{D}}_k\mathbf{D}_k^{-1/2}\|$. Notice that the matrix $\mathbf{D}_k^{-1/2}\hat{\mathbf{D}}_k\mathbf{D}_k^{-1/2}$ is block diagonal where its ith diagonal block is given by

$$\begin{aligned}&\mathbf{I} + \nabla_{\boldsymbol{y}^i\boldsymbol{y}^i} g(\boldsymbol{y}_k^i, \boldsymbol{y}_k^j; t_k)^{-\frac{1}{2}}\nabla_{\boldsymbol{y}^i\boldsymbol{y}^i} f^i(\boldsymbol{y}_k^i; t_k) \\ &\quad \times \nabla_{\boldsymbol{y}^i\boldsymbol{y}^i} g(\boldsymbol{y}_k^i, \boldsymbol{y}_k^j; t_k)^{-\frac{1}{2}}. \quad (59)\end{aligned}$$

Using the bounds in Assumptions 1 and 2, and the fact that for positive definite matrices $\lambda_{\min}(\boldsymbol{AB}) \geqslant \lambda_{\min}(\boldsymbol{A})\lambda_{\min}(\boldsymbol{B})$, we obtain that the eigenvalues of the matrix $\mathbf{D}_k^{-1/2}\hat{\mathbf{D}}_k\mathbf{D}_k^{-1/2}$ blocks are bounded below by $1 + 2m/L$. Thus, we obtain

$$\|\hat{\mathbf{D}}_k^{-1}\mathbf{D}_k\| = \|\mathbf{D}_k^{-1/2}\hat{\mathbf{D}}_k\mathbf{D}_k^{-1/2}\| \geqslant 1 + (2m/L)\,. \quad (60)$$

Since the eigenvalues of $\mathbf{D}_k^{-1/2}\hat{\mathbf{D}}_k\mathbf{D}_k^{-1/2}$ are lower bounded by $1 + 2m/L$, we obtain that

$$\|\mathbf{D}_k^{-1}\hat{\mathbf{D}}_k\| = \|\mathbf{D}_k^{1/2}\hat{\mathbf{D}}_k^{-1}\mathbf{D}_k^{1/2}\| \leqslant (1 + (2m/L))^{-1}\,. \quad (61)$$

[2]This conclusion depends on the condition number: large condition numbers favor Newton methods, small ones gradient methods, see also [10].

Substituting the upper bounds in (58) and (61) into (54) yields

$$\|\mathbf{D}_k^{-1/2}\mathbf{B}_k\mathbf{D}_k^{-1/2}\| \leqslant \varrho\,. \tag{62}$$

Note that (62) implies that the eigenvalues of $\mathbf{D}_k^{-1/2}\mathbf{B}_k\mathbf{D}_k^{-1/2}$ are strictly less than one, and thus, the expansion in (18) is valid.

We use the result in (62) to prove the claim in (42). Given the approximation in (19), we know that

$$\begin{aligned}\|\mathbf{H}_{k,(K)}^{-1}\| &\leqslant \|\mathbf{D}_k^{-1}\| \sum_{\tau=0}^{K} \|\mathbf{D}_k^{-1/2}\mathbf{B}_k\mathbf{D}_k^{-1/2}\|^{\tau} \\ &\leqslant \frac{1}{m+\ell/2}\sum_{\tau=0}^{K}\varrho^{\tau} = \frac{1-\varrho^{K+1}}{(m+\ell/2)(1-\varrho)}\end{aligned} \tag{63}$$

where the second inequality comes from formula for a finite geometric series. Moreover, we can derive an upper bound for the RHS of (63), and use the definition of ϱ to obtain

$$\frac{1-\varrho^{K+1}}{(m+\ell/2)(1-\varrho)} \leqslant \frac{1}{(m+\ell/2)(1-\varrho)} = \frac{2m+L}{m(2m+\ell)} =: H. \tag{64}$$

Combining the inequalities in (63) and (64), the claim in (42) follows. Moreover, the bound on the error e_k follows from [24, Proposition 3] with the definition of ϱ [cf., (62)], yielding

$$e_k = \|\mathbf{I} - \nabla_{\boldsymbol{yy}}F(\boldsymbol{y}_k;t_k)\mathbf{H}_{k,(K)}^{-1}\| = \|\mathbf{D}_k^{-\frac{1}{2}}\mathbf{B}_k\mathbf{D}_k^{-\frac{1}{2}}\|^{K+1} \tag{65}$$

from which the bound (43) follows. ■

Appendix B
Proof of Theorem 1

We only study the proof for DPC-G; the one for DAPC-G is similar (only the numerical constants change) and omitted in reason of space (the complete proof is available in [37, Appendix C]).

First, we establish that discrete-time sampling error bound stated in (48) is achieved by the updates of DPC-G. For simplicity, we modify the notation to omit the arguments $\boldsymbol{y}_k$ and t_k of the function F. In particular, define

$$\begin{aligned}&\nabla_{\boldsymbol{yy}}F := \nabla_{\boldsymbol{yy}}F(\boldsymbol{y}_k;t_k)\,, \quad \nabla_{t\boldsymbol{y}}F := \nabla_{t\boldsymbol{y}}F(\boldsymbol{y}_k;t_k)\\ &\nabla_{\boldsymbol{yy}}F^* := \nabla_{\boldsymbol{yy}}F(\boldsymbol{y}^*(t_k);t_k)\,,\ \nabla_{t\boldsymbol{y}}F^* := \nabla_{t\boldsymbol{y}}F(\boldsymbol{y}^*(t_k);t_k).\end{aligned} \tag{66}$$

Begin by considering the update of DPC-G, the prediction step, evaluated at a generic point $\boldsymbol{y}_k$ sampled at the current sample time t_k and with associated optimizer $\boldsymbol{y}^*(t)$

$$\boldsymbol{y}^*(t_{k+1}) = \boldsymbol{y}^*(t_k) - h\,[\nabla_{\boldsymbol{yy}}F^*]^{-1}\nabla_{t\boldsymbol{y}}F^* + \boldsymbol{\Delta}_k. \tag{67}$$

Rewrite the approximate prediction step $\boldsymbol{y}_{k+1|k} = \boldsymbol{y}_k + h\,\boldsymbol{p}_k$ by adding and subtracting the exact prediction step $h\,[\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}F$, yielding

$$\begin{aligned}\boldsymbol{y}_{k+1|k} &= \boldsymbol{y}_k + h\boldsymbol{p}_{k,(K)} \\ &\quad + h\,[\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}F - h\,[\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}F.\end{aligned} \tag{68}$$

Subtract (67) from (68), take the norm, and apply the triangle inequality to the resulted expression to obtain

$$\begin{aligned}\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\| &\leqslant \|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + \\ &\quad h\,\left\|[\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}\,F - [\nabla_{\boldsymbol{yy}}F^*]^{-1}\nabla_{t\boldsymbol{y}}F^*\right\| \\ &\quad + h\,\left\|\boldsymbol{p}_{k,(K)} - [\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}F\right\| + \|\boldsymbol{\Delta}_k\|.\end{aligned} \tag{69}$$

We proceed to analyze the three terms on the RHS of (69). The last term $\|\boldsymbol{\Delta}_k\|$ is bounded above by $h^2\Delta$ as in (41).

We proceed to find an upper bound for the second summand in the RHS of (69). We use the same reasoning as in [10, Appendix B], which yields (cf. [10, Eq. (62)])

$$\begin{aligned}h\,\left\|[\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}F - [\nabla_{\boldsymbol{yy}}F^*]^{-1}\nabla_{t\boldsymbol{y}}F^*\right\| \leqslant \\ \frac{C_0C_1h}{m^2}\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + \frac{C_2h}{m}\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\|.\end{aligned} \tag{70}$$

Finally, we proceed to analyze the third term in (69). Rewrite this term using the definition of the prediction step $\boldsymbol{p}_{k,(K)} = -\mathbf{H}_{k,(K)}^{-1}\nabla_{t\boldsymbol{y}}F$, and apply the mixed first-order partial derivative bound $\|\nabla_{t\boldsymbol{y}}F(\boldsymbol{y};t)\| \leqslant C_0$ stated in Assumption 2 to obtain

$$\begin{aligned}h\,\|\mathbf{H}_{k,(K)}^{-1}\nabla_{t\boldsymbol{y}}F - \nabla_{\boldsymbol{yy}}F^{-1}\nabla_{t\boldsymbol{y}}F\| \leqslant \\ C_0h\,\|\mathbf{H}_{k,(K)}^{-1} - \nabla_{\boldsymbol{yy}}F^{-1}\|\,.\end{aligned} \tag{71}$$

Observe that $\|\mathbf{H}_{k,(K)}^{-1} - \nabla_{\boldsymbol{yy}}F^{-1}\|$ is bounded above by $\|\nabla_{\boldsymbol{yy}}F^{-1}\|\|\nabla_{\boldsymbol{yy}}F\mathbf{H}_{k,(K)}^{-1} - \mathbf{I}\|$. This observation in conjunction with the upper bound for the error vector $e_k = \|\nabla_{\boldsymbol{yy}}F\mathbf{H}_{k,(K)}^{-1} - \mathbf{I}\|$ in Proposition 3 implies that

$$\begin{aligned}\|\mathbf{H}_{k,(K)}^{-1} - \nabla_{\boldsymbol{yy}}F^{-1}\| \leqslant \\ \|\nabla_{\boldsymbol{yy}}F^{-1}\|\|\nabla_{\boldsymbol{yy}}F\mathbf{H}_{k,(K)}^{-1} - \mathbf{I}\| \leqslant \frac{\varrho^{K+1}}{m}.\end{aligned} \tag{72}$$

Combining the results in (71) and (72) shows that the third in the RHS of (69) is upper bounded by

$$h\,\left\|\boldsymbol{p}_{k,(K)} - [\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}F\right\| \leqslant \frac{hC_0}{m}\varrho^{K+1}\,. \tag{73}$$

By substituting the bounds in (70) and (73) into (69) and considering the definitions of σ in (44) and $\Gamma(\varrho,K)$ in Theorem 1, we obtain

$$\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \sigma\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + h\,\Gamma(\varrho,K) + h^2\Delta. \tag{74}$$

For the correction step [cf., (29)] , we may use the standard property of projected gradient descent for strongly convex functions with Lipschitz gradients. The Euclidean error norm of the projected gradient descent method converges linearly as

$$\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \rho\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\| \tag{75}$$

where $\rho = \max\{|1-\gamma m|, |1-\gamma(L+M)|\}$; see e.g., [10]. Plug the correction error in (75) into the prediction error in (74) to obtain

$$\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \rho\sigma\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + \rho\,\varphi \tag{76}$$

where $\varphi := h\,\Gamma(\varrho, K) + h^2\Delta$. Therefore

$$\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*(t_{k+1})\| \leqslant (\rho\sigma)^{k+1}\|\boldsymbol{y}_0 - \boldsymbol{y}^*(t_0)\| + \rho\,\varphi\sum_{i=0}^{k}(\rho\sigma)^i. \tag{77}$$

Substitute $k+1$ by k and simplify the sum in (77), making use of the fact that $\rho\sigma < 1$, which yields

$$\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| \leqslant (\rho\sigma)^k\|\boldsymbol{y}_0 - \boldsymbol{y}^*(t_0)\| + \rho\,\varphi\left[\frac{1-(\rho\sigma)^k}{1-\rho\sigma}\right]. \tag{78}$$

Observing (78) together with the definition of φ, (48) follows. In particular, if $\rho\sigma < 1$ (that is (47) holds), the sequence $\{\boldsymbol{y}_k\}$ converges Q-linearly to $\boldsymbol{y}^*$ up to an error bound as

$$\limsup_{k\to\infty}\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| = \left[\frac{\rho}{1-\rho\sigma}\right]\left(h\,\Gamma(\varrho,K) + h^2\left[\frac{C_0C_2}{m^2} + \frac{C_3}{2m} + \frac{C_0^2C_1}{2m^3}\right]\right). \tag{79}$$

To establish the result stated in (46), observe that in the worst case, we may upper bound the term $\|[\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}F - [\nabla_{\boldsymbol{yy}}F^*]^{-1}\nabla_{t\boldsymbol{y}}F^*\|$ in (69) by $\|[\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}F\| + \|[\nabla_{\boldsymbol{yy}}F^*]^{-1}\nabla_{t\boldsymbol{y}}F^*\|$, which yields

$$\|[\nabla_{\boldsymbol{yy}}F]^{-1}\nabla_{t\boldsymbol{y}}F - [\nabla_{\boldsymbol{yy}}F^*]^{-1}\nabla_{t\boldsymbol{y}}F^*\| \leqslant \frac{2C_0}{m}. \tag{80}$$

Substituting the upper bound in (80) into (69) yields

$$\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + h\,\frac{2C_0}{m} + \varphi. \tag{81}$$

Using the definition $\varphi = h\,\Gamma(\varrho,K) + h^2\Delta$ and observing the relation in (75), we can write

$$\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \rho\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + \rho\left[2h\frac{C_0}{m} + \varphi\right]. \tag{82}$$

Recursively applying (82) backwards to the initial time

$$\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| \leqslant \rho^k\|\boldsymbol{y}_0 - \boldsymbol{y}^*(t_0)\| + \rho\left[\frac{2hC_0}{m} + \varphi\right]\left[\frac{1-\rho^k}{1-\rho}\right] \tag{83}$$

and by sending $k\to\infty$, we can simplify (83) as

$$\limsup_{k\to\infty}\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| = \frac{2C_0\rho h}{m(1-\rho)} + \frac{\rho}{1-\rho}h\Gamma(\varrho,K) + h^2\frac{\rho}{1-\rho}\left[\frac{C_0C_2}{m^2} + \frac{C_3}{2m} + \frac{C_0^2C_1}{2m^3}\right] \tag{84}$$

which is (46). The result in (84) holds if $\rho < 1$, which is the case if $\gamma < 2/(L+M)$, as stated in Theorem 1. ■

Appendix C
Proof of Theorem 2

Also in this case, we prove the theorem only for DPC-N, while for DAPC-N—which is very similar—is given in [37, Appendix E].

Since DPC-N and DPC-G are identical in their prediction steps, we may consider the prediction error result established during the proof Theorem 1, i.e., the expression in (74) with $k=0$. We turn our attention to the correction step, and consider, in particular, the gap to the optimal trajectory before and after correction at time t_{k+1} as

$$\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*(t_{k+1})\| = \|\boldsymbol{y}_{k+1|k} - \gamma\,\mathbf{H}_{k+1|k}^{-1}\nabla_{\boldsymbol{y}}F - \boldsymbol{y}^*(t_{k+1})\|. \tag{85}$$

Subsequently, we simplify notation by defining the shorthands

$$\begin{aligned}
&\nabla_{\boldsymbol{y}}F := \nabla_{\boldsymbol{y}}F(\boldsymbol{y}_{k+1|k}; t_{k+1})\ ,\ \nabla_{\boldsymbol{yy}}F := \nabla_{\boldsymbol{yy}}F(\boldsymbol{y}_{k+1|k}; t_{k+1})\\
&\nabla_{\boldsymbol{y}}F^* := \nabla_{\boldsymbol{y}}F(\boldsymbol{y}^*(t_{k+1}); t_{k+1}),\\
&\nabla_{\boldsymbol{yy}}F^* := \nabla_{\boldsymbol{yy}}F(\boldsymbol{y}^*(t_{k+1}); t_{k+1})\\
&\nabla_{t\boldsymbol{y}}F := \nabla_{t\boldsymbol{y}}F(\boldsymbol{y}_{k+1|k}; t_{k+1})\ ,\\
&\nabla_{t\boldsymbol{y}}F^* := \nabla_{t\boldsymbol{y}}F(\boldsymbol{y}^*(t_{k+1}); t_{k+1})\ .
\end{aligned} \tag{86}$$

Add and subtract $\gamma\,\nabla_{\boldsymbol{yy}}F^{-1}\nabla_{\boldsymbol{y}}F$ to the expression inside the norm in (85) which is the exact damped Newton, and use the triangle inequality to obtain

$$\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1}) - \gamma\nabla_{\boldsymbol{yy}}F^{-1}\nabla_{\boldsymbol{y}}F\| + \gamma\|(\nabla_{\boldsymbol{yy}}F^{-1} - \mathbf{H}_{k+1|k}^{-1})\nabla_{\boldsymbol{y}}F\|. \tag{87}$$

We proceed to bound the two terms in the RHS of (87). Consider the first term: left multiply by $\nabla_{\boldsymbol{yy}}F$ and its inverse, and left factor out the Hessian inverse $\nabla_{\boldsymbol{yy}}F^{-1}$. Making use of the Cauchy–Schwartz inequality, the first term of RHS of (87) is bounded above as

$$\begin{aligned}
\|\boldsymbol{y}_{k+1|k} - \gamma\nabla_{\boldsymbol{yy}}F^{-1}\nabla_{\boldsymbol{y}}F - \boldsymbol{y}^*(t_{k+1})\| &\leqslant (1-\gamma)\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\|\\
&+ \gamma\|\nabla_{\boldsymbol{yy}}F^{-1}\|\|\nabla_{\boldsymbol{yy}}F(\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})) - \nabla_{\boldsymbol{y}}F\|\ .
\end{aligned} \tag{88}$$

We use now the same arguments as in [10, Appendix C, Eqs. (83)–(85)] to show that (88) can be upper bounded by

$$(1-\gamma)\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\| + \frac{\gamma C_1}{2m}\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\|^2\ . \tag{89}$$

With this upper estimate in place for the first term on the RHS of (87), we shift focus to the second term. Use the Cauchy–Schwartz inequality to obtain $\|(\nabla_{\boldsymbol{yy}}F^{-1} - \mathbf{H}_{k+1|k}^{-1})\nabla_{\boldsymbol{y}}F\|$ is bounded above by $\|\nabla_{\boldsymbol{yy}}F^{-1}\|\|\nabla_{\boldsymbol{yy}}F\mathbf{H}_{k+1|k}^{-1} - \mathbf{I}\|\|\nabla_{\boldsymbol{y}}F\|$. Use the upper bound $1/m$ for the spectrum of $\nabla_{\boldsymbol{yy}}F^{-1}$ and the Hessian approximation error [cf., (43)] to obtain

$$\|(\nabla_{\boldsymbol{yy}}F^{-1} - \mathbf{H}_{k+1|k}^{-1})\nabla_{\boldsymbol{y}}F\| \leqslant \frac{\varrho^{K'+1}}{m}\|\nabla_{\boldsymbol{y}}F\|. \tag{90}$$

Now, focusing on the second term in the product on the RHS of (90), we use of the optimality criterion of $\boldsymbol{y}^*(t_{k+1})$, which is equivalent to $\nabla_{\boldsymbol{y}}F^* = \mathbf{0}$, to write

$$\|\nabla_{\boldsymbol{y}}F\| = \|\nabla_{\boldsymbol{y}}F - \nabla_{\boldsymbol{y}}F^*\| \leqslant (L+M)\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\| \tag{91}$$

where we have used the Lipschitz property of the gradients. Substituting the upper bound in (91) into (90) and considering the definition $\Gamma(\varrho, K') = (C_0/m)\varrho^{K'+1}$ lead to

$$\|(\nabla_{\boldsymbol{yy}} F^{-1} - \mathbf{H}_{k+1|k}^{-1})\nabla_{\boldsymbol{y}} F\| \leqslant \frac{L+M}{C_0}\Gamma(\varrho, K')\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}(t_{k+1})\|. \quad (92)$$

Apply the bounds (89)–(92) to the RHS of (87)

$$\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \gamma\frac{C_1}{2m}\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\|^2 + \left(\gamma\frac{L+M}{C_0}\Gamma(\varrho, K')+1-\gamma\right)\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\|. \quad (93)$$

Now we consider the prediction step, which by (74), we have

$$\|\boldsymbol{y}_{k+1|k} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \sigma\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + \varphi \quad (94)$$

with $\varphi = h\,\Gamma(\varrho, K) + h^2\Delta$ as defined in Appendix B. Substituting the relation (94) into (93) allows us to write

$$\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \gamma\frac{C_1}{2m}\left(\sigma\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + \varphi\right)^2 + \left(\gamma\frac{L+M}{C_0}\Gamma(\varrho, K') + 1 - \gamma\right)\left(\sigma\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + \varphi\right). \quad (95)$$

The RHS of (95) is a quadratic function of the error $\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\|$ at time t_k, which upper bounds the error sequence at the subsequent time t_{k+1}. For certain selections of parameters K, K', and h, (95) defines a contraction. To determine the conditions for which this occurs, we solve for an appropriate radius of contraction. Let $\tau > 0$ be a positive scalar such that

$$\alpha_2\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\|^2 + \alpha_1\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + \alpha_0 \leqslant \tau\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| + \alpha_0 \quad (96)$$

where the coefficients α_0, α_1, and α_2 of the quadratic polynomial of the error $\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\|$ are defined from the RHS of (95), and given as

$$\alpha_2 = \gamma\frac{C_1}{2m}\sigma^2,\ \alpha_1 = \sigma\left[\gamma\frac{C_1}{m}\varphi + \gamma\frac{L+M}{C_0}\Gamma(\varrho, K') + 1 - \gamma\right]$$
$$\alpha_0 = \varphi\left[\gamma\frac{C_1}{2m}\varphi + \gamma\frac{L+M}{C_0}\Gamma(\varrho, K') + 1 - \gamma\right]. \quad (97)$$

Based on (96), to guarantee the Q-linear convergence of the error sequence $\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\|$, we require $\tau < 1$, which by simple algebra it is satisfied if

$$\alpha_1 < \tau, \quad \|\boldsymbol{y}_0 - \boldsymbol{y}^*(t_0)\| \leqslant (\tau - \alpha_1)/\alpha_2\ . \quad (98)$$

Finally, we need to require that the second condition in (98) holds true for all k, that is $\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*(t_{k+1})\| \leqslant \|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\|$, which implies

$$\tau(\tau - \alpha_1)/\alpha_2 + \alpha_0 \leqslant (\tau - \alpha_1)/\alpha_2. \quad (99)$$

Conditions (98) and (99), the definitions of α_0, α_1, and α_2 in (97), and $\varphi = h\,\Gamma(\varrho, K) + h^2\Delta$ establish the small enough conditions on sampling period and optimality gap as well as the large enough conditions on the approximation levels K, K' in Theorem 2, for any chosen $\tau < 1$. In particular, the terms α_0 and α_1 are polynomial functions of the sampling period h and the approximation levels K and K'. Conditions (98) and (99) describe a system of nonlinear inequalities for any fixed $1 - \gamma < \tau < 1$. For arbitrarily small h and large K and K', α_0 and α_1 can be made 0 and $1 - \gamma$, respectively. Formally

$$\lim_{h\to 0,\, K, K'\to\infty}\alpha_0 = 0, \quad \lim_{h\to 0,\, K, K'\to\infty}\alpha_1 = 1 - \gamma. \quad (100)$$

When $\alpha_0 = 0$ and $\alpha_1 = 1 - \gamma$ conditions (98) and (99) hold with attraction region $\bar{R} = 2m(\tau - 1 + \gamma)/\gamma C_1\sigma^2$. Since α_0 and α_1 go to their limits monotonically with h, K, K', then—by continuity—there exists a large enough $\bar{K}$ and small enough attraction region $\bar{R}$ for which conditions (98) and (99) can be achieved. The convergence region in (98) is dictated by the Newton step and by the step-size choice γ. If the cost function is a time-varying quadratic function, then $C_1 = 0$, and the convergence region is the whole space. If γ is chosen very small, then also in this case the convergence region becomes arbitrarily large, as expected.

By selecting the error polynomial coefficients as (97) and recursively applying (96) backward in time, we obtain

$$\|\boldsymbol{y}_k - \boldsymbol{y}^*(t_k)\| \leqslant \tau^k\|\boldsymbol{y}_0 - \boldsymbol{y}^*(t_0)\| + \alpha_0\left[\frac{1-\tau^k}{1-\tau}\right] \quad (101)$$

which since $\tau < 1$, the right side of the (101) is finite, implying (49), after the expansion of the coefficients. ■

REFERENCES

[1] A. Simonetto, A. Mokhtari, A. Koppel, G. Leus, and A. Ribeiro, "A decentralized prediction-correction method for networked time-varying convex optimization," in *Proc. IEEE 6th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process.*, Cancun, Mexico, Dec. 2015, pp. 509–512.

[2] A. Simonetto, A. Koppel, A. Mokhtari, G. Leus, and A. Ribeiro, "A quasi-newton prediction-correction method for decentralized dynamic convex optimization," in *Proc. Eur. Control Conf.*, Aalborg, Denmark, Jun. 2016, pp. 1934–1939.

[3] P. Alriksson and A. Rantzer, "Distributed Kalman filter using weighted averaging," in *Proc. 17th Int. Symp. Math. Theory Netw. Syst.*, Kyoto, Japan, Jul. 2006, pp. 1–6.

[4] M. Farina, G. Ferrari-Trecate, and R. Scattolini, "Distributed moving horizon estimation for linear constrained systems," *IEEE Trans. Autom. Control*, vol. 55, no. 11, pp. 2462–2475, Nov. 2010.

[5] P. Ögren, E. Fiorelli, and N. Leonard, "Cooperative control of mobile sensor networks: Adaptive gradient climbing in a distributed environment," *IEEE Trans. Autom. Control*, vol. 49, no. 8, pp. 1292–1302, Aug. 2004.

[6] F. Borrelli and T. Keviczky, "Distributed LQR design for identical dynamically decoupled systems," *IEEE Trans. Autom. Control*, vol. 53, no. 8, pp. 1901–1912, Sep. 2008.

[7] F. Arrichiello, "Coordination control of multiple mobile robots," Ph.D. thesis, Università degli studi di Cassino, Cassino, Italy, 2006.

[8] Y. Kim and M. Mesbahi, "On maximizing the second smallest eigenvalue of a state-dependent graph Laplacian," *IEEE Trans. Autom. Control*, vol. 51, no. 1, pp. 116–120, Jan. 2006.

[9] R. Graham and J. Cortes, "Adaptive information collection by robotic sensor networks for spatial estimation," *IEEE Trans. Autom. Control*, vol. 57, no. 6, pp. 1404–1419, Jun. 2012.

[10] A. Simonetto, A. Mokhtari, A. Koppel, G. Leus, and A. Ribeiro, "A class of prediction-correction methods for time-varying convex optimization," *IEEE Trans. Signal Process.*, vol. 64, no. 17, pp. 4576–4591, Sep. 2016.

[11] B. T. Polyak, *Introduction to Optimization*. Optimization Software, Inc., New York, NY, USA, 1987.

[12] V. M. Zavala and M. Anitescu, "Real-time nonlinear optimization as a generalized equation," *SIAM J. Control Optim.*, vol. 48, no. 8, pp. 5444–5467, 2010.

[13] A. L. Dontchev, M. I. Krastanov, R. T. Rockafellar, and V. M. Veliov, "An Euler-Newton continuation method for tracking solution trajectories of parametric variational inequalities," *SIAM J. Control Optim.*, vol. 51, no. 51, pp. 1823–1840, 2013.
[14] M. Kamgarpour and C. Tomlin, "Convergence properties of a decentralized Kalman filter," in *Proc. 47th IEEE Conf. Decis. Control*, Cancun, Mexico, Dec. 2008, pp. 3205–3210.
[15] S.-Y. Tu and A. H. Sayed, "Mobile adaptive networks," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 649–664, Aug. 2011.
[16] D. Bajovic, D. Jakovetic, J. Xavier, B. Sinopoli, and J. M. F. Moura, "Distributed detection via Gaussian running consensus: Large deviations asymptotic analysis," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4381–4396, Sep. 2011.
[17] M. M. Zavlanos, A. Ribeiro, and G. J. Pappas, "Network integrity in mobile robotic networks," *IEEE Trans. Autom. Control*, vol. 58, no. 1, pp. 3–18, Jan. 2013.
[18] F. Y. Jakubiec and A. Ribeiro, "D-MAP: Distributed maximum a posteriori probability estimation of dynamic systems," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 450–466, Jan. 15, 2013.
[19] Q. Ling and A. Ribeiro, "Decentralized dynamic optimization through the alternating direction method of multipliers," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1185–1197, Mar. 1, 2014.
[20] A. Simonetto and G. Leus, "Distributed asynchronous time-varying constrained optimization," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2014, pp. 2142–2146.
[21] A. Koppel, A. Simonetto, A. Mokhtari, G. Leus, and A. Ribeiro, "Target tracking with dynamic convex optimization," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2015, pp. 1210–1214.
[22] H. Yin, P. Mehta, S. Meyn, and U. V. Shanbhag, "Learning in mean-field games," *IEEE Trans. Autom. Control*, vol. 59, no. 3, pp. 629–644, Mar. 2014.
[23] P. Guan, M. Raginsky, and R. Willett, "Online Markov decision processes with Kullback-Leibler control cost," *IEEE Trans. Autom. Control*, vol. 59, no. 6, pp. 1423–1438, Jun. 2014.
[24] A. Mokhtari, Q. Ling, and A. Ribeiro, "Network Newton—Part I: Algorithm and convergence," arXiv preprint arXiv:1504.06017, 2015.
[25] A. Mokhtari, Q. Ling, and A. Ribeiro, "Network Newton—Part II: Convergence rate and implementation," arXiv preprint arXiv:1504.06020, 2015.
[26] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.
[27] A. Loukas, M. A. Zúñiga, I. Protonotarios, and J. Gao, "How to identify global trends from local decisions? event region detection on mobile networks," in *Proc. Int. Conf. Comput. Commun.*, Toronto, CA, USA, Apr. 2014, pp. 1177–1185.
[28] V. Roy, S. Gishkori, and G. Leus, "Spatial rainfall mapping from path-averaged rainfall measurements exploiting sparsity," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Altanta, GA, USA, Dec. 2014, pp. 321–325.
[29] J. A. Martinez, R. Heusdens, and R. C. Hendriks, "A generalized fourier domain: Signal processing framework and applications," *Signal Process.*, vol. 93, no. 5, pp. 1259–1267, 2013.
[30] B. Grocholsky, A. Makarenko, and H. Durrant-Whyte, "Information-theoretic coordinated control of multiple sensor platforms," in *Proc. IEEE Int. Conf. Robot. Autom.*, Taipei, Taiwan, Sep. 2003, pp. 1521–1526.
[31] N. E. Leonard, D. A. Paley, R. E. Davis, D. M. Fratantoni, F. Lekien, and F. Zhang, "Coordinated control of an underwater glider fleet in an adaptive ocean sampling field experiment in monterey bay," *J. Field Robot.*, vol. 27, no. 6, pp. 718–740, 2010.
[32] L. Xiao and S. Boyd, "Optimal scaling of a gradient method for distributed resource allocation," *J. Optim. Theory. Appl.*, vol. 129, no. 3, pp. 469–488, 2006.
[33] E. Ghadimi, I. Shames, and M. Johansson, "Multi-step gradient methods for networked optimization," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5417–5429, Nov. 2013.
[34] P. Wan and M. D. Lemmon, "Event-triggered distributed optimization in sensor networks," in *Proc. Int. Conf. Inf. Process. Sensor Netw.*, San Francisco, CA, USA, Apr. 2009, pp. 49–60.
[35] D. G. Feingold and R. S. Varga, "Block diagonally dominant matrices and generalization of the Gerschgorin circle theorem," *Pacific J. Math.*, vol. 12, no. 4, pp. 1241–1250, 1962.
[36] A. L. Dontchev and R. T. Rockafellar, *Implicit Functions and Solution Mappings*. New Yoek, NY, USA: Springer, 2009.
[37] A. Simonetto, A. Koppel, A. Mokhtari, G. Leus, and A. Ribeiro, "Decentralized prediction-correction methods for networked time-varying convex optimization," arXiv:1602.01716, 2016.
[38] E. K. Ryu and S. Boyd, "Primer on monotone operator methods," *Appl. Comput. Math.*, vol. 15, pp. 3–43, 2016.

Andrea Simonetto (M'12) received the Ph.D. degree in systems and control from the Delft University of Technology, Delft, The Netherlands, in 2012.

He is a Research Staff Member with the Optimization and Control Group, IBM Research Ireland, Dublin, Ireland. His current research interests include optimization and control for large scale systems, with applications in smart energy and transportation.

Alec Koppel received the B.A. degree in mathematics and the M.S. degree in systems science from Washington University, St. Louis, MO, USA, in 2011 and 2012, respectively. He is currently working toward the Ph.D. degree with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA.

He is a participant in the SMART Scholarship program sponsored by the U.S. Army Research Laboratory, Adelphi, MD, USA. His research focuses on signal processing, machine learning, and optimization.

Aryan Mokhtari received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2011, and the M.Sc. degree in electrical engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2014. Since 2012, he has been working toward the Ph.D. degree in the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA.

His research interests include stochastic and distributed optimization and machine learning.

Geert Leus (F'12) received the M.Sc. and Ph.D. degrees in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in June 1996 and May 2000, respectively.

He is currently a Full Professor with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands.

Dr. Leus received the 2002 IEEE Signal Processing Society Young Author Best Paper Award and the 2005 IEEE Signal Processing Society Best Paper Award. He is a Fellow of the European Association for Signal Processing. He was the Chair of the IEEE Signal Processing for Communications and Networking Technical Committee. He is currently a Member-at-Large to the Board of Governors of the IEEE Signal Processing Society and he serves as the Editor in Chief of the *EURASIP Journal on Advances in Signal Processing*.

Alejandro Ribeiro (M'07) received the B.Sc. degree in electrical engineering from the Universidad de la Republica Oriental del Uruguay, Montevideo, Uruguay, in 1998 and the M.Sc. and Ph.D. degree in electrical engineering from the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA, in 2005 and 2007, respectively.

From 1998 to 2003, he was a member of the Technical Staff with Bellsouth, Montevideo. After his M.Sc. and Ph.D studies, in 2008, he joined the University of Pennsylvania, Philadelphia, PA, USA, where he is currently the Rosenbluth Associate Professor with the Department of Electrical and Systems Engineering.