Distributed Dual Subgradient Methods with Averaging and Applications to Grid Optimization

Haitian Liu[1] · **Subhonmesh Bose[2]** · **Hoa Dinh Nguyen[3]** · **Ye Guo[1]** · **Thinh T. Doan[4]** · **Carolyn L. Beck[2]**

Abstract

We study finite-time performance of a recently proposed distributed dual subgradient (DDSG) method for convex-constrained multi-agent optimization problems. The algorithm enjoys performance guarantees on the last primal iterate, as opposed to those derived for ergodic means for standard DDSG algorithms. Our work improves the recently published convergence rate of $\mathcal{O}(\log T / \sqrt{T})$ with decaying step-sizes to $\mathcal{O}(1/\sqrt{T})$ with constant step-size on a metric that combines sub-optimality and constraint violation. We then numerically evaluate the algorithm on three grid optimization problems. Namely, these are tie-line scheduling in multi-area power systems, coordination of distributed energy resources in radial distribution networks, and joint dispatch of transmission and distribution assets. The DDSG algorithm applies to each problem with various relaxations and linearizations of the power flow equations. The numerical experiments illustrate various properties of the DDSG algorithm–comparison with standard DDSG, impact of the number of agents, and why Nesterov-style acceleration can fail in DDSG settings.

Keywords Distributed optimization · Power system examples

1 Introduction

Distributed optimization algorithms offer mechanisms to optimize in multi-agent environments, where one cannot aggregate all problem data in a central location. Agents in this paradigm iteratively perform local computational steps and communicate relevant variables over a network. While a variety of distributed solution architectures

---

✉ Subhonmesh Bose
  boses@illinois.edu

Extended author information available on the last page of the article

have been developed and analyzed in the literature for various settings, we focus on distributed dual subgradient (DDSG) methods with averaging that can be used to solve convex-constrained multi-agent optimization problems of the form

$$\mathcal{P} : \underset{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}{\text{minimize}} \qquad \sum_{j=1}^{N} f_j(\boldsymbol{x}_j), \tag{1a}$$

$$\text{subject to} \qquad \sum_{j=1}^{N} \boldsymbol{g}_j^{\mathrm{E}}(\boldsymbol{x}_j) = 0, \tag{1b}$$

$$\sum_{j=1}^{N} \boldsymbol{g}_j^{\mathrm{I}}(\boldsymbol{x}_j) \leq 0, \tag{1c}$$

$$\boldsymbol{x}_j \in \mathbb{X}_j \subseteq \mathbb{R}^{n_j}, \ j = 1,\dots,N. \tag{1d}$$

The N agents communicate only across edges of an undirected graph $\mathfrak{G}(N, \mathbb{E})$. Assume that functions f_j, $\boldsymbol{g}_j^{\mathrm{I}}$ are convex and $\boldsymbol{g}_j^{\mathrm{E}}$ is affine over the compact convex set $\mathbb{X}_j$ for each $j = 1,\dots,N$.

At its core, DDSG methods rely on dual decomposition that starts by separating the Lagrangian into agent-wise Lagrangian functions that each agent optimizes, given a dual iterate (see classical texts such as [7, 61]). This agent-wise optimization of the primal variables can be shown to provide a subgradient of the dual function at the dual iterate. Thus, a subgradient ascent on the dual function can be used to solve the dual problem. Such an update rule requires a central coordinator to manage the dual iterates, which is undesirable in many distributed contexts. DDSG methods circumvent this need by maintaining *local copies* of such multipliers and running a consensus-based distributed dual ascent on these local multiplier copies. Approximate primal solutions can be recovered from these dual solutions as in [63], building on techniques in [31, 42, 50, 55], among others; asymptotic guarantees on recovered primal sequences are also known.

In this paper, we study a distributed dual subgradient method, analyzed recently in [48], that provides a fully distributed variant of the algorithm, also proposed recently in [57]. As opposed to vanilla (i.e., standard) DDSG algorithms, convergence guarantees of these algorithms are obtained for the *last primal iterates* as opposed to that for *ergodic means* of primal sequences. As Nesterov and Shikhman argue in [57], iterates in vanilla DDSG algorithms can oscillate wildly during algorithm execution. If the intermediate primal iterates of the algorithm are implemented in practice, vanilla DDSG can negatively impact stability in multi-agent environments. Besides, primal recovery is no longer required for asymptotic constraint satisfaction in this paradigm.

For the algorithms in [48, 57], asymptotic guarantees for the convergence of the last primal-dual iterate to the set of primal-dual optimizers have been established with decaying step-sizes. Also, the authors of [48] characterize a $\mathcal{O}(\log T / \sqrt{T})$-bound on a metric that combines sub-optimality and constraint violation with decaying step-sizes (similar to [20]), while they argue a lower bound of $\mathcal{O}(1/\sqrt{T})$. Their analysis shows an $\mathcal{O}(1/\sqrt{T})$-bound, when a coordinator manages the dual updates centrally.

Thus, they view the $\log T$ factor as a price for decentralization. Our first contribution of this paper is to close this gap in Sect. 2 with the proof in Sect. 3, i.e., we sharpen the convergence rate to $\mathcal{O}(1/\sqrt{T})$ for this fully distributed algorithm with a *constant step-size*. The dual function for $\mathcal{P}$ can be nonsmooth, and as a result, a first-order algorithm that climbs the dual function cannot converge at a better rate (see Appendix A), making the algorithm order-optimal.

A number of algorithms apply to $\mathcal{P}$ and its variants. Dual (sub)gradient algorithms include our key motivators in [48, 57], and [22, 58] with accelerated counterparts discussed in [28, 37]. As opposed to using optimal solutions of primal subproblems to update the dual variables, one can incrementally update both primal and dual variables via saddle-point dynamics as in [34, 47, 52]. With specific constraint types in multi-agent settings, distributing dual variable updates similar to the algorithm we study has been considered in [62]. Bundle methods have been investigated in [15, 65] that leverage lower and upper estimates of dual functions and their subdifferential sets using the history of inexact evaluations of the same. While a detailed performance comparison of each of these algorithms with the DDSG algorithm we study in our paper remains out of scope, these prior works feature in several comparative remarks we make in the sequel.

Throughout the rest of the paper, we study various properties of our DDSG algorithm on $\mathcal{P}$. Specifically, we consider three different optimization problems that arise in operations of the electric power system and cast them as examples of $\mathcal{P}$. For each problem, we adopt different power flow models, different notions of agents and different definitions of the network over which these agents interact. This exercise stands as our second contribution, which demonstrates the modeling power of $\mathcal{P}$ and the broad applicability of the algorithm with convergence guarantees. The grid optimization problems are nonconvex due to the nature of Kirchhoff's laws [12]. In this paper, we *convexify* each problem by considering approximations and convex relaxations for power flow models that are suited to that application. Each problem and the application of DDSG illustrates specific properties of the algorithm that are delineated below.

- *Multi-area optimal power flow* ($\mathcal{P}_1$) seeks to dispatch resources over an interconnected transmission network, parts of which are controlled by different system operators. The distributed algorithmic architecture we study in Sect. 4 bypasses the need for the system operators to share all relevant data from within their footprint with another system operator, and yet seeks to solve a joint optimal power flow problem through a distributed solution architecture, e.g., see [30]. For $\mathcal{P}_1$, we consider a linear power flow model for transmission networks motivated by the use of such models in wholesale market environments that facilitate inter-area coordination, e.g., in [64]. Through the numerical example on $\mathcal{P}_1$, we demonstrate how the averaging scheme stabilizes the last iterate, compared to vanilla DDSG methods, and circumvents the need for primal recovery. We also show that the finite-time constraint violation of the algorithm is empirically better than its theoretical bound suggests.
- *Coordination of distributed energy resources (DERs) in distribution grids* ($\mathcal{P}_2$) is designed to optimize real and reactive power outputs from DERs at the grid-edge to minimize cost (disutility) of such a dispatch and possibly a network-wide objec-

tive such as frequency regulation. A distributed algorithm allows quick updates of optimization variables without the need to communicate with a central coordinator across the distribution grid. For $\mathcal{P}_2$ in Sect. 5, we consider a second-order cone programming (SOCP)-based relaxation of the power flow equations in the distribution grids. The thoroughly studied SOCP-based relaxation of power flow equations for distribution grids are often tight in practice (see [23, 25]). Through examples of $\mathcal{P}_2$, we show that our algorithm can track changing grid conditions for small networks. Tracking performance degrades with network size, where aggregation of nodes as agents becomes vital for performance. Stated differently, if speed is paramount, one must carefully control the degree of decentralization, possibly opting for an alternate distributed solution architecture with a coordinator, or devising a fully centralized algorithm.

- *Transmission and distribution (T&D) grid coordination* ($\mathcal{P}_3$) seeks to dispatch assets across the transmission and distribution grids without the need to collect all information from the grid-edge and the bulk power systems at one location. The distributed solution architecture in Sect. 6 alleviates the transmission system operator's lack of visibility into utility-managed distribution networks. For $\mathcal{P}_3$, we consider two different power flow models for the transmission and the distribution grids. For transmission, we choose a semidefinite programming (SDP)-based relaxation of power flow equations, given its popularity to approach the AC optimal power flow problem [9, 43, 68]. For the distribution grids, we consider the linear distribution power flow model from [5]. Using an example of $\mathcal{P}_3$, we compare the empirical performance of the DDSG method with an adopted distributed acceleration scheme proposed in [28], where acceleration fails to increase convergence speed with linear costs, but performs well with quadratic costs. We argue that nonsmoothness of the dual function with linear costs is behind this phenomenon, and elaborate on its role in Appendix A.

We recognize that a long literature has emerged on each of the grid optimization problems; we provide a non-exhaustive list in Table 1. Our goal in presenting the power system examples is not an attempt to empirically compare our DDSG algorithm with others in the literature proposed for these problems. Rather, our formulations of these problems as instances of $\mathcal{P}$ reveal the generality of $\mathcal{P}$ and make it possible for subsequent unified algorithm development for $\mathcal{P}$. In addition, each example highlights an important aspect of the DDSG algorithm.

2 The Dual Subgradient Method with Averaging

We present the DDSG algorithm to solve $\mathcal{P}$ in (1). The finite-time performance guarantee for $\mathcal{P}$ is derived in Sect. 3. Then in Sects. 4, 5 and 6, we cast $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ as examples of $\mathcal{P}$.

To lighten notation, let $\boldsymbol{g}_j$ collect both $\boldsymbol{g}_j^{\mathrm{E}}$ and $\boldsymbol{g}_j^{\mathrm{I}}$ with the understanding that the first M^{E} constraints encode equalities and the last M^{I} are inequalities. The algorithm relies on Lagrangian duality theory associated with $\mathcal{P}$. We begin by defining the Lagrangian

function,

$$\mathcal{L}(\boldsymbol{x}, z) := \sum_{j=1}^{N} \left[f_j(\boldsymbol{x}_j) + z^{\mathsf{T}} \boldsymbol{g}_j(\boldsymbol{x}_j) \right], \tag{2}$$

$$\boldsymbol{x}^{\mathsf{T}} = (\boldsymbol{x}_1^{\mathsf{T}}, \dots, \boldsymbol{x}_N^{T}) \in \mathbb{X} := \mathbb{X}_1 \times \dots \times \mathbb{X}_N,$$

$$z \in \mathbb{Z} := \mathbb{R}^{M^{\mathrm{E}}} \times \mathbb{R}_+^{M^{\mathrm{I}}}. \tag{3}$$

Then, $\mathcal{P}$ can be cast as a min-max problem with optimal value $\mathcal{P}^\star$ as

$$\mathcal{P}^\star = \min_{\boldsymbol{x} \in \mathbb{X}} \max_{z \in \mathbb{Z}} \ \mathcal{L}(\boldsymbol{x}, z). \tag{4}$$

Let $\mathbb{X}^\star$ denote the set of optimizers of $\mathcal{P}$. Associated with $\mathcal{P}$ is its dual problem

$$\mathcal{P}_D^\star = \max_{z \in \mathbb{Z}} \min_{\boldsymbol{x} \in \mathbb{X}} \ \mathcal{L}(\boldsymbol{x}, z). \tag{5}$$

Let $\mathbb{Z}^\star$ denote the set of optimizers of the dual problem. Weak duality implies that $\mathcal{P}^\star \geq \mathcal{P}_D^\star$. We say strong duality holds if the inequality is met with an equality. Further $\boldsymbol{x}^\star \in \mathbb{X}, z^\star \in \mathbb{Z}$ is a saddle point of $\mathcal{L}$, if

$$\mathcal{L}(\boldsymbol{x}^\star, z) \leq \mathcal{L}(\boldsymbol{x}^\star, z^\star) \leq \mathcal{L}(\boldsymbol{x}, z^\star), \tag{6}$$

for all $\boldsymbol{x} \in \mathbb{X}, z \in \mathbb{Z}$. The well-known saddle point theorem (see [24, Theorem 2.156]) states that the primal-dual optimizers $\mathbb{X}^\star \times \mathbb{Z}^\star$ coincide with the saddle points of $\mathcal{L}$. We assume throughout that the set of saddle points of $\mathcal{P}$ is nonempty and bounded. As a result, strong duality holds for $\mathcal{P}$, i.e., $\mathcal{P}^\star = \mathcal{P}_D^\star$, and the set of primal-dual optimizers is nonempty. Saddle-points exist under standard constraint qualifications such as Slater's condition, e.g., see [24, Theorem 2.165].

Dual decomposition techniques for distributed optimization rely on the observation that the dual function separates into agent-wise optimization problems, given a multiplier z as

$$\min_{\boldsymbol{x} \in \mathbb{X}} \ \mathcal{L}(\boldsymbol{x}, z) = \sum_{j=1}^{N} \underbrace{\min_{\boldsymbol{x}_j \in \mathbb{X}_j} \mathcal{L}_j(\boldsymbol{x}_j, z)}_{:= \mathcal{D}_j(z)}, \tag{7}$$

where $\mathcal{L}_j(\boldsymbol{x}_j, z) := f_j(\boldsymbol{x}_j) + z^{\mathsf{T}} \boldsymbol{g}_j(\boldsymbol{x}_j)$. If the agents can perform these agent-wise minimizations, then a distributed projected subgradient ascent algorithm can solve the dual problem (e.g., see [10]). Per Danskin's theorem, a subgradient $\nabla_z \mathcal{D}_j(z)$ can be obtained from the agent-wise minimization of $\mathcal{L}_j$, given that the subdifferential set of the concave function $\mathcal{D}_j$ at z is

$$\partial_z \mathcal{D}_j(z) := \mathrm{conv}\{\partial_z \mathcal{L}_j(\boldsymbol{x}_j, z) \mid \boldsymbol{x}_j \in \mathbb{X}_j^\star(z)\}. \tag{8}$$

Here, "conv" computes the convex hull of its argument and $\mathbb{X}_j^\star(z)$ is the set of minimizers of $\mathcal{L}_j(\cdot, z)$ over $\mathbb{X}_j$. The minimization problem is well-defined, given that $\mathbb{X}_j$'s are compact. Running such an algorithm, however, requires a *central coordinator* to compute the z-update and broadcast the results to all agents.

To avoid coordination for the dual update, one can alternately create *local copies* of z's among all agents and enforce equality among these local estimates in the dual problem as

$$\max \sum_{j=1}^{N} \mathcal{D}_j(z_j), \text{ subject to } z_j = z_k, \ j, k = 1, \ldots, N, \tag{9}$$

where z_j is the local copy of z with agent j. One can run a projected distributed subgradient ascent as in [63] to solve (9). The primal iterates obtained from agent-wise minimization of $\mathcal{L}_j$ evaluated at the dual iterates may fail to collectively satisfy the constraints of $\mathcal{P}$. Primal averaging schemes have been studied in [63]; limit points of such recovered primal solutions are known to satisfy the constraints. One can judiciously maintain local copies only among a subset of the agents to relieve communication burden (see [38]).

Recently, a dual subgradient algorithm was proposed in [57] that leveraged an estimation sequence technique to provide guarantees on sub-optimality and infeasibility on the *last iterate*. This algorithm does not treat ergodic means simply as outputs from a dual subgradient calculation, but rather uses these means as primal-dual iterates to run the algorithm. We focus on the fully distributed variant of the algorithm that is proposed and analyzed in [48]. To present the algorithm, let $W \in \mathbb{R}^{N \times N}$ be a doubly stochastic, irreducible, and aperiodic weighting matrix that follows the sparsity pattern of $\mathfrak{G}$, i.e.,

$$W_{j,k} \neq 0 \iff (j, k) \in \mathbb{E}. \tag{10}$$

Then, the distributed projected dual subgradient with averaging is given by Algorithm 1, where x_j/X_j are primal sequences and z_j/Z_j are dual sequences. The updates comprise minimization of the local dual function in step 3, averaging of these primal minimizers in step 4, a consensus followed by local subgradient-based dual update in step 5, and an ergodic mean computation for the projected dual variable in step 6 with step-size η. Here, $\pi_\mathbb{Z}$ projects the arguments on $\mathbb{Z}$.

We remark that bundle methods from [15] with its distributed variant in [65] can be used to solve $\mathcal{P}$. Algorithm 1 shares parallels with these bundle methods in that z/Z carry *memory* of past subgradient and function evaluations–or their inexact variants–in its updates rather than classic dual subgradient algorithms that only process local subgradient information in each dual update. However, bundle methods are somewhat qualitatively different as they leverage the execution history of the algorithm to approximate the dual function to generate successive iterates. Understanding deeper connections between these two methods in how they exploit the geometry of the dual function defines an interesting direction for future research.

Algorithm 1: Distributed dual subgradient with averaging to solve $\mathcal{P}$.

1 Choose $z_j(1) = 0$, $Z_j(0) = 0$, $x_j(0) \in \mathbb{X}_j$ and $\eta = \eta_0/\sqrt{T}$.

2 **for** $t = 1, \ldots, T$ **do**

3      $X_j(t) \leftarrow \operatorname{argmin}_{x_j \in \mathbb{X}_j} \mathcal{L}_j(x_j, z_j(t))$.

4      $x_j(t) \leftarrow \frac{t-1}{t} x_j(t-1) + \frac{1}{t} X_j(t)$.

5      $Z_j(t) \leftarrow \sum_{k=1}^{N} W_{jk} Z_k(t-1) + t g_j(x_j(t)) - (t-1) g_j(x_j(t-1))$.

6      $z_j(t+1) \leftarrow \frac{t}{t+1} z_j(t) + \frac{1}{t+1} \pi_{\mathbb{Z}} \left[\eta Z_j(t) \right]$.

7 **end**

To study convergence properties of this algorithm, consider the metric introduced in [57] and used in [48], given by

$$\mathcal{V}_T(\boldsymbol{x}(T), \overline{z}(T)) := \sum_{j=1}^{N} f_j(\boldsymbol{x}_j(T)) - \sum_{j=1}^{N} \mathcal{D}_j(\overline{z}(T)) + \frac{\eta T}{2N} \left\| \pi_{\mathbb{Z}} \left[\sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(T)) \right] \right\|^2, \tag{11}$$

where $\overline{z}(T) := \frac{1}{T} \sum_{t=1}^{T} \pi_{\mathbb{Z}} \left[\eta \overline{Z}(t-1) \right]$, $\overline{Z}(t) := \frac{1}{N} \sum_{j=1}^{N} Z_j(t)$. The sum of the first two terms measures the gap between the primal objective at $\boldsymbol{x}(T) \in \mathbb{X}$ and the dual function evaluated at $\overline{z}(T) \in \mathbb{Z}$. The last summand is a measure of the constraint violation at $\boldsymbol{x}(T)$. We sharpen the bound of [48, Theorem 2] in the next result.

Theorem 1 *Iterates generated by Algorithm 1 with $\eta = \eta_0/\sqrt{T}$ over $t = 1, \ldots, T$, $\eta_0 > 0$ constant, satisfy*

$$\mathcal{V}_T(\boldsymbol{x}(T), \overline{z}(T)) \leq \frac{1}{\sqrt{T}} \left(\frac{C_0}{1 - \sigma_2(\boldsymbol{W})} + C_1 \right), \tag{12a}$$

$$\mathcal{V}_T(\boldsymbol{x}(T), \overline{z}(T)) \geq \mathcal{P}^\star - \sum_{j=1}^{N} \mathcal{D}_j(\overline{z}(T)) - \frac{C_2}{\sqrt{T}}. \tag{12b}$$

where C's are positive constants that do not depend on $\mathfrak{G}$ or T, and $\sigma_2(\boldsymbol{W})$ is the second largest singular value of $\boldsymbol{W}$.

Our upper bound in this result sharpens the conclusion of [48, Theorem 2], while the lower bound is identical. The result implies that the metric in (11) indeed converges at a rate of $1/\sqrt{T}$. Our proof of the bounds largely mirrors that of [48, Theorem 2], but deviates from the reliance on results from [20] that incur the $\log T$ factor. Instead, we use an argument inspired by the proof of [17, Theorem 2].

We briefly remark on the implication of Theorem 1 on the sub-optimality of $\boldsymbol{x}(T)$ and the constraint violation, separately. Call the right hand side of the upper bound in (12) as $C'/\sqrt{T}$. Then, we infer

$$\sum_{j=1}^{N} f_j(\boldsymbol{x}_j(T)) - \mathcal{P}^\star \leq \sum_{j=1}^{N} f_j(\boldsymbol{x}_j(T)) - \sum_{j=1}^{N} \mathcal{D}_j(\overline{z}(T)) \leq \frac{C'}{\sqrt{T}}, \tag{13}$$

since $\mathcal{P}^\star$ dominates the dual function, and the second summand of $\mathscr{V}_T$ in (11) is non-negative. Also, combining the two inequalities in (12), we get

$$\mathcal{P}^\star - \sum_{j=1}^{N} \mathcal{D}_j(\overline{z}(T)) \leq \frac{C' + C_2}{\sqrt{T}}. \tag{14}$$

Thus, the last primal and the dual iterate exhibit an $\mathcal{O}(1/\sqrt{T})$ sub-optimality. The dual function can be nonsmooth at an optimum. This convergence rate is therefore order-optimal, given [56, Theorem 3.2.1]. Faster convergence guarantees through a Nesterov-style acceleration require stronger assumptions. See Sect. 6 and Appendix A for a discussion.

Providing the same rate for constraint violation using Theorem 1 remains challenging. The difficulty stems from the fact that, unless $x(T)$ is feasible in $\mathcal{P}$, the primal-dual gap can assume negative values. However, this gap is bounded below. Using (13), we obtain

$$\sum_{j=1}^{N} f_j(x_j(T)) - \sum_{j=1}^{N} \mathcal{D}_j(\overline{z}(T)) \geq \min_{x \in \mathbb{X}} \sum_{j=1}^{N} f_j(x_j) - \max_{x \in \mathbb{X}} \sum_{j=1}^{N} f_j(x_j) =: -D_f. \tag{15}$$

The constant $D_f \geq 0$ is finite, owing to the compact nature of $\mathbb{X}$. Then, (12) implies

$$\frac{\eta_0}{2N} \left\| \pi_{\mathbb{Z}} \left[\sum_{j=1}^{N} g_j(x(T)) \right] \right\|^2 \leq \frac{D_f}{\sqrt{T}} + \frac{C'}{T}. \tag{16}$$

This suggests a worst-case $\mathcal{O}(T^{-1/4})$ decay in constraint violation–an estimate that is overly conservative as our numerical estimates will reveal. Better guarantees for vanilla DDSG methods are known, e.g., in [54, 55]. A constant step-size of $\eta_0/\sqrt{T}$ yields an $\mathcal{O}(1/\sqrt{T})$ convergence of the ergodic mean of the primal iterates.

With non-summable and square-summable decaying step-sizes, vanilla DDSG methods converge to a *single* dual optimizer (not just to the optimal set), even in distributed settings, e.g., see [31]. While asymptotic convergence to the primal-dual optimal set for Algorithm 1 is established in [48, Lemmas 1, 2], convergence of the dual iterates to a single dual optimizer has not been established. It remains unclear whether such a result is attainable beyond vanilla DDSG; the conclusion does not hold even for centralized first-order primal-dual methods; see [51, Sect. 3.2].

3 Proof of Theorem 1

We begin by defining additional notation. Since f_j and g_j's are convex (and hence, continuous) and $\mathbb{X}_j$ is compact for each j, these functions admit positive constants

$D_{\mathbb{X}}$, D_g, L_g such that

$$\left\| x_j - x_j' \right\| \le D_{\mathbb{X}}, \quad \left\| g_j(x_j) \right\| \le D_g, \quad \left\| g_j(x_j) - g_j(x_j') \right\| \le L_g \left\| x_j - x_j' \right\| \tag{17}$$

for all $x_j, x_j' \in \mathbb{X}_j$. Also, we define $D_Z := L_g D_{\mathbb{X}} + D_g$.

3.1 Bounding $\mathscr{V}_T$ from Above

Using this notation, we derive the upper bound on $\mathscr{V}_T$ in four steps:

(a) We bound the duality gap as

$$\sum_{j=1}^{N} \left[f_j(x_j(T)) - \mathcal{D}_j(\overline{z}(T)) \right] \le \frac{2D_g}{T} \sum_{j=1}^{N} \sum_{t=1}^{T} \eta \left\| Z_j(t-1) - \overline{Z}(t-1) \right\|$$
$$- \frac{\eta}{T} \sum_{j=1}^{N} \sum_{t=1}^{T} g_j(X_j(t))^{\mathsf{T}} \pi_{\mathbb{Z}} \left[\overline{Z}(t-1) \right]. \tag{18}$$

(b) Then, we bound the constraint violation as

$$\frac{T}{2N} \left\| \pi_{\mathbb{Z}} \left[\sum_{j=1}^{N} g_j(x(T)) \right] \right\|^2 \le \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} g_j\left(X_j(t) \right)^{\mathsf{T}} \pi_{\mathbb{Z}} \left[\overline{Z}(t-1) \right] + \frac{1}{2} N D_Z^2. \tag{19}$$

(c) We prove that Z_j's remain close to their centroid as

$$\sum_{j=1}^{N} \left\| Z_j(t) - \overline{Z}(t) \right\|_2 \le N^{3/2} D_Z \left(1 - \sigma_2(W) \right)^{-1}. \tag{20}$$

(d) Steps (a), (b), (c) are combined to prove the result.

• *Step (a). Bounding the duality gap:* Note that

$$\sum_{j=1}^{N} \left[f_j(x_j(T)) - \mathcal{D}_j(\overline{z}(T)) \right] = \sum_{j=1}^{N} \left[f_j(x_j(T)) - \mathcal{D}_j(z_j(T)) \right]$$
$$+ \sum_{j=1}^{N} \left[\mathcal{D}_j(z_j(T)) - \mathcal{D}_j(\overline{z}(T)) \right]$$

$$\leq \sum_{j=1}^{N} \underbrace{\left[\frac{1}{T} \sum_{t=1}^{T} f_j \left(X_j(t) \right) - \mathcal{D}_j(z_j(T)) \right]}_{:= \mathscr{A}_j}$$

$$+ D_g \sum_{j=1}^{N} \underbrace{\left\| z_j(T) - \overline{z}(T) \right\|}_{:= \mathscr{B}_j}. \tag{21}$$

The last line follows from three observations: f_j is convex, $x_j(T) = \frac{1}{T} \sum_{t=1}^{T} X_j(t)$ and $\mathcal{D}_j$ is D_g-Lipschitz. In the rest of step (a), we individually bound $\mathscr{A}_j$ and $\mathscr{B}_j$.

To obtain a bound on $\mathscr{A}_j$, note that

$$t z_j(t) - (t-1) z_j(t-1) = \pi_{\mathbb{Z}} \left[\eta Z_j(t-1) \right], \tag{22}$$

which then implies

$$\begin{aligned} t \mathcal{L}_j(X_j(t), z_j(t)) &= \mathcal{L}_j(X_j(t), t z_j(t) - (t-1) z_j(t-1)) \\ &\quad + (t-1) \mathcal{L}_j(X_j(t), z_j(t-1)) \\ &\geq \mathcal{L}_j \left(X_j(t), \pi_{\mathbb{Z}} \left[\eta Z_j(t-1) \right] \right) \\ &\quad + (t-1) \mathcal{L}_j(X_j(t-1), z_j(t-1)). \end{aligned} \tag{23}$$

The first line follows from elementary algebra, while the second line requires the definition of Z_j and the fact that $X_j(t-1)$ minimizes $\mathcal{L}_j(\cdot, z_j(t-1))$ over $\mathbb{X}_j$. Iterating the above inequality, we obtain

$$T \mathcal{D}_j(z_j(T)) = T \mathcal{L}_j(X_j(T), z_j(T)) \geq \sum_{t=1}^{T} \mathcal{L}_j \left(X_j(t), \pi_{\mathbb{Z}} \left[\eta Z_j(t-1) \right] \right). \tag{24}$$

The above relation bounds $\mathscr{A}_j$ from above as

$$\begin{aligned} \mathscr{A}_j &\leq \frac{1}{T} \sum_{t=1}^{T} \left[f_j \left(X_j(t) \right) - \mathcal{L}_j \left(X_j(t), \pi_{\mathbb{Z}} \left[\eta Z_j(t-1) \right] \right) \right] \\ &= -\frac{1}{T} \sum_{t=1}^{T} g_j(X_j(t))^{\mathsf{T}} \pi_{\mathbb{Z}} \left[\eta Z_j(t-1) \right] \\ &= -\frac{1}{T} \sum_{t=1}^{T} g_j(X_j(t))^{\mathsf{T}} \left(\pi_{\mathbb{Z}} \left[\eta Z_j(t-1) \right] - \pi_{\mathbb{Z}} \left[\eta \overline{Z}(t-1) \right] \right) \\ &\quad - \frac{1}{T} \sum_{t=1}^{T} g_j(X_j(t))^{\mathsf{T}} \pi_{\mathbb{Z}} \left[\eta \overline{Z}(t-1) \right]. \end{aligned} \tag{25}$$

The Cauchy-Schwarz inequality and the bounded nature of $\boldsymbol{g}_j$ imply

$$\mathscr{A}_j \le \frac{D_g}{T} \sum_{t=1}^{T} \left\| \pi_{\mathbb{Z}} \left[\eta \boldsymbol{Z}_j(t-1) \right] - \pi_{\mathbb{Z}} \left[\eta \overline{\boldsymbol{Z}}(t-1) \right] \right\|$$
$$- \frac{\eta}{T} \sum_{t=1}^{T} \boldsymbol{g}_j(\boldsymbol{X}_j(t))^{\mathsf{T}} \pi_{\mathbb{Z}} \left[\overline{\boldsymbol{Z}}(t-1) \right]. \tag{26}$$

To bound $\mathscr{B}_j$, we use the definition of $z_j(t)$ to infer

$$z_j(T) = \frac{1}{T} \sum_{t=1}^{T} \pi_{\mathbb{Z}} \left[\eta \boldsymbol{Z}_j(t-1) \right], \tag{27}$$

which in turn implies

$$\mathscr{B}_j \le \frac{1}{T} \sum_{t=1}^{T} \left\| \pi_{\mathbb{Z}} \left[\eta \boldsymbol{Z}_j(t-1) \right] - \pi_{\mathbb{Z}} \left[\eta \overline{\boldsymbol{Z}}(t-1) \right] \right\|. \tag{28}$$

Using the bounds of (26) and (28) in (21) and appealing to the non-expansive nature of the projection operator yields (18), completing step (a) of the proof.

• *Step (b). Bounding the constraint violation:* From the $\boldsymbol{Z}$-update, we obtain

$$\overline{\boldsymbol{Z}}(t) = \frac{t}{N} \sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(t)), \tag{29}$$

which proves useful in bounding the constraint violation as

$$\frac{T^2}{N^2} \left\| \pi_{\mathbb{Z}} \left[\sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(T)) \right] \right\|^2 = \left\| \pi_{\mathbb{Z}} \left[\overline{\boldsymbol{Z}}(T) \right] \right\|^2$$
$$= \sum_{t=1}^{T} \left(\left\| \pi_{\mathbb{Z}} \left[\overline{\boldsymbol{Z}}(t) \right] \right\|^2 - \left\| \pi_{\mathbb{Z}} \left[\overline{\boldsymbol{Z}}(t-1) \right] \right\|^2 \right)$$
$$\le 2 \sum_{t=1}^{T} \underbrace{\left[\pi_{\mathbb{Z}} \left[\overline{\boldsymbol{Z}}(t-1) \right]^{\mathsf{T}} \left[\overline{\boldsymbol{Z}}(t) - \overline{\boldsymbol{Z}}(t-1) \right] \right]}_{:=\mathscr{E}(t)}$$
$$+ \sum_{t=1}^{T} \underbrace{\left\| \overline{\boldsymbol{Z}}(t) - \overline{\boldsymbol{Z}}(t-1) \right\|^2}_{:=\mathscr{F}(t)}. \tag{30}$$

The inequality follows from the fact that for any two scalars a, b, we have

$$a^2 - b^2 = 2b(a - b) + (a - b)^2,$$

$$\left(\pi_{\mathbb{R}_+}[a]\right)^2 - \left(\pi_{\mathbb{R}_+}[b]\right)^2 \le 2\pi_{\mathbb{R}_+}[b](a - b) + (a - b)^2. \tag{31}$$

We separately bound $\mathscr{E}(t)$ and $\mathscr{F}(t)$. For the former, we use the convexity of $\boldsymbol{g}_j$ and the $\boldsymbol{x}$-update to conclude

$$
\begin{aligned}
\overline{\boldsymbol{Z}}(t) - \overline{\boldsymbol{Z}}(t - 1) &= \frac{t}{N} \sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(t)) - \frac{t-1}{N} \sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(t - 1)) \\
&= \frac{t}{N} \sum_{j=1}^{N} \boldsymbol{g}_j \left(\frac{t-1}{t} \boldsymbol{x}_j(t - 1) + \frac{1}{t} \boldsymbol{X}_j(t) \right) \\
&\quad - \frac{t-1}{N} \sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(t - 1)) \\
&\le \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{g}_j \left(\boldsymbol{X}_j(t) \right).
\end{aligned}
\tag{32}
$$

Note that if an entry of $\boldsymbol{g}$ encodes an equality constraint, the linearity of that constraint implies that the above relation is met with an equality. Thus, we obtain

$$\mathscr{E}(t) \le \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{g}_j \left(\boldsymbol{X}_j(t) \right)^{\mathsf{T}} \pi_{\mathbb{Z}} \left[\overline{\boldsymbol{Z}}(t - 1) \right]. \tag{33}$$

To bound $\mathscr{F}(t)$, we use the first line of (32) and the bounded/Lipschitz nature of $\boldsymbol{g}_j$ on $\mathbb{X}_j$ to get

$$
\begin{aligned}
\left\| \overline{\boldsymbol{Z}}(t) - \overline{\boldsymbol{Z}}(t - 1) \right\| &\le \frac{L_g}{N} \sum_{j=1}^{N} (t - 1) \left\| \boldsymbol{x}_j(t) - \boldsymbol{x}_j(t - 1) \right\| + D_g \\
&= \frac{L_g}{N} \sum_{j=1}^{N} \left\| \boldsymbol{X}_j(t) - \boldsymbol{x}_j(t) \right\| + D_g \\
&\le L_g D_{\mathbb{X}} + D_g \\
&= D_{\mathbb{Z}}.
\end{aligned}
\tag{34}
$$

Thus, we have $\mathscr{F}(t) \le D_{\mathbb{Z}}^2$. Replacing the bounds on $\mathscr{E}(t)$ and $\mathscr{F}(t)$ in (30) gives the required bound on constraint violation in (19), completing the proof of step (b).

● *Step (c): Bounding the deviation of Z_j's from their centroid:* Consider $\zeta \in \mathbb{R}^{N \times M}$, given by

$$\zeta(t)^{\mathsf{T}} = \left(Z_1(t) \mid \dots \mid Z_N(t) \right). \tag{35}$$

and define $\Delta := I - \frac{1}{N} \mathbb{1}\mathbb{1}^{\mathsf{T}}$, where $\mathbb{1} \in \mathbb{R}^N$ is a vector of all ones and $I \in \mathbb{R}^{N \times N}$ is the identity matrix. Using this notation, we deduce

$$\sum_{j=1}^{N} \left\| Z_j(t) - \overline{Z}(t) \right\|_2 \leq \sqrt{N} \left\| \Delta \zeta(t) \right\|_F \leq N \left\| \Delta \zeta(t) \right\|_2, \tag{36}$$

where, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Then, the Z-updates can be written as

$$\zeta(t+1) = W \zeta(t) + \varphi(t), \quad \zeta(0) = 0 \tag{37}$$

with $\varphi(t) \in \mathbb{R}^{N \times M}$; an analysis similar to (34) gives that each row has a 2-norm bounded above by D_Z, implying

$$\|\varphi(t)\|_2 \leq \sqrt{N} D_Z. \tag{38}$$

Using (37), we then obtain

$$\|\Delta \zeta(t+1)\|_2 = \|\Delta \left(W \zeta(t) + \varphi(t) \right)\|_2 \leq \|W \Delta \zeta(t)\|_2 + \|\Delta \varphi(t)\|_2 \tag{39}$$

utilizing the fact that W and Δ commute. To bound the first term in (39), note that W is doubly stochastic for which the Perron-Frobenius theorem [33, Theorem 8.4.4] implies that its eigenvalue with the largest absolute value is unity for which $\mathbb{1}$ is the eigenvector. However, $\mathbb{1}^{\mathsf{T}} \Delta = 0$, which in turn suggests $\Delta \zeta(t)$ is orthogonal to this eigenvector. Using the Courant-Fischer theorem [33, Theorem 4.2.6], we then obtain

$$\|W \Delta \zeta(t)\|_2 \leq \sigma_2(W) \|\Delta \zeta(t)\|_2, \tag{40}$$

where $\sigma_2(W)$ is the second largest singular value of W. Since W is irreducible and aperiodic, $\sigma_2(W) \in (0, 1)$. We bound the second term in (39) as

$$\|\Delta \varphi(t)\|_2 \leq \underbrace{\|\Delta\|_2}_{=1} \|\varphi(t)\|_2 \leq \sqrt{N} D_Z, \tag{41}$$

because the 2-norm is sub-multiplicative. Using the bounds in (40) and (41) in (39), imply

$$\|\Delta \zeta(t+1)\|_2 \leq \sigma_2(W) \|\Delta \zeta(t)\|_2 + \sqrt{N} D_Z. \tag{42}$$

Iterating the above inequality gives

$$\|\mathbf{\Delta}\boldsymbol{\zeta}(t)\|_2 \leq \sqrt{N} D_Z \sum_{\ell=0}^{t-1} [\sigma_2(\boldsymbol{W})]^{t-\ell-1} \leq \sqrt{N} D_Z (1 - \sigma_2(\boldsymbol{W}))^{-1}. \quad (43)$$

Then, (36) and (43) imply (20), finishing step (c) of the proof.

• *Step (d). Combining steps (a), (b), (c) to derive the result:* Note that (18) and (19) together with the definition of $\mathcal{V}_T$ give

$$
\begin{aligned}
\mathcal{V}_T(\boldsymbol{x}(T), \overline{\boldsymbol{z}}(T) &= \sum_{j=1}^{N} \left[f_j(\boldsymbol{x}_j(T)) - \mathcal{D}_j(\overline{\boldsymbol{z}}(T)) \right] + \frac{\eta T}{2N} \left\| \pi_{\mathbb{Z}} \left[\sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(T)) \right] \right\|^2 \\
&\leq \frac{2 D_g}{T} \sum_{j=1}^{N} \sum_{t=1}^{T} \eta \left\| \boldsymbol{Z}_j(t-1) - \overline{\boldsymbol{Z}}(t-1) \right\| + \frac{\eta}{2} N D_Z^2 \\
&\leq \eta \frac{2 D_g N^{3/2} D_Z}{1 - \sigma_2(\boldsymbol{W})} + \frac{\eta}{2} N D_Z^2.
\end{aligned}
\quad (44)
$$

where the second inequality follows from (20). Using $\eta = \eta_0/\sqrt{T}$, we then obtain the upper bound in (12).

3.2 Bounding $\mathcal{V}_T$ from Below

By the saddle-point property of a primal-dual optimizer $(\boldsymbol{x}^\star, \boldsymbol{z}^\star)$ of $\mathcal{P}$, we get

$$
\begin{aligned}
\mathcal{P}^\star &= \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{z}^\star) \\
&\leq \mathcal{L}(\boldsymbol{x}(T), \boldsymbol{z}^\star) \\
&= \sum_{j=1}^{N} f_j(\boldsymbol{x}_j(T)) + \boldsymbol{z}^{\star,\mathsf{T}} \sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(T)) \\
&\leq \sum_{j=1}^{N} f_j(\boldsymbol{x}_j(T)) + \boldsymbol{z}^{\star,\mathsf{T}} \pi_{\mathbb{Z}} \left[\sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(T)) \right].
\end{aligned}
\quad (45)
$$

Applying Young's inequality to the last summand in the right hand side of the above relation, we further get

$$\mathcal{P}^\star \leq \sum_{j=1}^{N} f_j(\boldsymbol{x}_j(T)) + \frac{N}{2\eta T} \|\boldsymbol{z}^\star\|^2 + \frac{\eta T}{2N} \left\| \pi_{\mathbb{Z}} \left[\sum_{j=1}^{N} \boldsymbol{g}_j(\boldsymbol{x}_j(T)) \right] \right\|^2. \quad (46)$$

Subtracting $\sum_{j=1}^{N} \mathcal{D}_j(\overline{z}(T))$ on both sides and using $\eta = \eta_0/\sqrt{T}$ yields the desired lower bound on $\mathcal{V}_T$ in (12).

To summarize, the lower and upper bound of $\mathcal{V}_T(\boldsymbol{x}(T), \overline{\boldsymbol{z}}(T))$ are

$$\mathcal{P}^{\star} - \sum_{j=1}^{N} \mathcal{D}_j(\overline{z}(T)) - \frac{N}{2\eta T} \left\| \boldsymbol{z}^{\star} \right\|^2 \leq \mathcal{V}_T \leq \eta \frac{2 D_g N^{3/2} D_Z}{1 - \sigma_2(\boldsymbol{W})} + \frac{\eta}{2} N D_Z^2, \qquad (47)$$

with $\eta = \eta_0/\sqrt{T}$. Note that D_g, D_Z and N impact the required iterations that depend on the problem structure. One expects D_g, and as a result, D_Z to grow with the number of constraints (1b), (1c), and N to grow with the number of agents.

4 Grid Optimization Example 1: Tie-Line Scheduling

In this section and the next two, we present three different examples of grid optimization problems that can be cast as examples of $\mathcal{P}$. Section 7 for a selective literature survey on these problems. Here, we focus instead on reformulation of each as an instance of $\mathcal{P}$ and report numerical results from running Algorithm 1 to reveal interesting properties of the algorithm.

We first present our results on $\mathcal{P}_1$–the tie-line scheduling problem. Tie-lines are transmission lines that interconnect the footprints of different system operators (henceforth, called areas). Ideally, one would solve a joint optimal power flow (OPF) problem over assets within all areas to optimize tie-line schedules. However, technical and legal challenges impede aggregation of all problem data at a central location, requiring a distributed algorithm design. Since the seminal work in [39], a substantial literature has developed on multi-area OPF problems for tie-line scheduling. Here, we formulate $\mathcal{P}_1$ and apply Algorithm 1.

Denote by $\boldsymbol{\theta}_j \in \mathbb{R}^{n_j}$ and $\boldsymbol{\theta}_{\overline{j}} \in \mathbb{R}^{n_{\overline{j}}}$, the voltage phase angles at the internal and the boundary buses in each area j, respectively. We adopt a linear power flow model in which the vector of power injections within an area (generation less demand $\boldsymbol{p}_j^G - \boldsymbol{p}_j^D$ at internal buses and zero injections at boundary buses) become linear in voltage phase angles through suitably defined matrices $\boldsymbol{B}$. Individual line flows within and across areas also become linear in these angles, defined through matrices $\boldsymbol{H}$. Angles are constrained within $[0, 2\pi]$, represented as $\boldsymbol{\theta} \in \Theta$. Utilizing $\boldsymbol{L}$'s to encode line capacities, c's to denote power procurement costs (typically deduced from supply offers in electricity markets), and $\mathfrak{G}(N, \mathbb{E})$ to represent the interconnection graph among the areas, the multi-area OPF problem becomes

$$\mathcal{P}_1 : \text{minimize} \qquad \sum_{j=1}^{N} c_j \left(\boldsymbol{p}_j^G \right),$$

$$\text{subject to} \qquad \underline{\boldsymbol{p}}_j^G \leq \boldsymbol{p}_j^G \leq \overline{\boldsymbol{p}}_j^G, \boldsymbol{\theta} \in \Theta, \qquad (48a)$$

$$\boldsymbol{B}_{j,j} \boldsymbol{\theta}_j + \boldsymbol{B}_{j,\overline{j}} \boldsymbol{\theta}_{\overline{j}} = \boldsymbol{p}_j^G - \boldsymbol{p}_j^D, \qquad (48b)$$

Fig. 1 The three-area network for multi-area optimal power flow simulations, obtained by joining three IEEE 118-bus systems
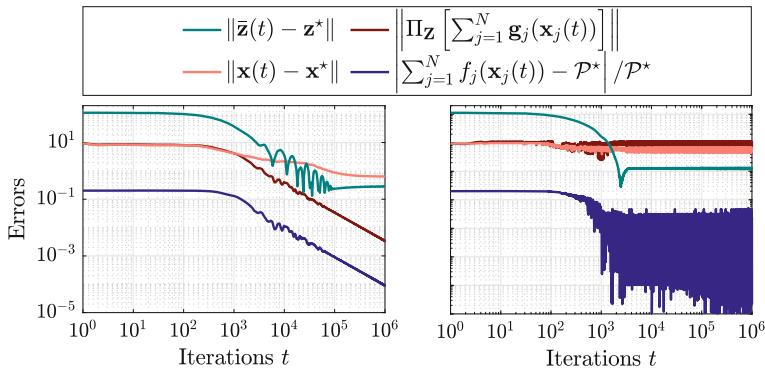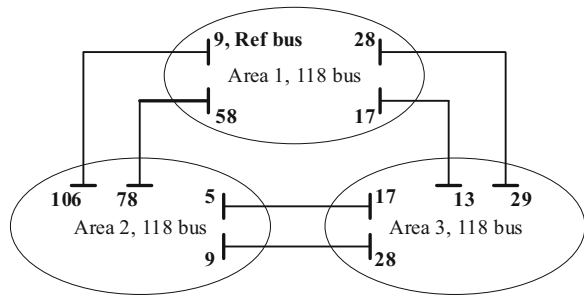
Fig. 2 Performance of Algorithm 1 (left) and Algorithm 2 (right) on $\mathcal{P}_1$ for the network in Fig. 1

$$B_{\bar{j},j}\boldsymbol{\theta}_j + B_{\bar{j},\bar{j}}\boldsymbol{\theta}_{\bar{j}} + \sum_{k\sim j} B_{\bar{j},\bar{k}}\boldsymbol{\theta}_{\bar{k}} = 0, \qquad (48c)$$

$$H_j\boldsymbol{\theta}_j + H_{\bar{j}}\boldsymbol{\theta}_{\bar{j}} \leq L_j, \qquad (48d)$$

$$H_{j,k}\boldsymbol{\theta}_{\bar{j}} + H_{k,j}\boldsymbol{\theta}_{\bar{k}} \leq L_{jk},$$

$$j = 1, \ldots, N, \ k \sim j \text{ in } \mathfrak{G}. \qquad (48e)$$

Here, (48a)–(48d) encode the generation capacity and angle constraints, power balance and transmission line constraints within each area, while (48e) enforces limits on tie-line flows. To cast (48) as $\mathcal{P}$, define

$$\boldsymbol{x}_j = \left(\boldsymbol{\theta}_j^{\mathsf{T}}, \boldsymbol{\theta}_{\bar{j}}^{\mathsf{T}}, [\boldsymbol{p}_j^G]^{\mathsf{T}}\right)^{\mathsf{T}}, \ \mathbb{X}_j = \left\{\boldsymbol{x}_j \mid (48a), (48b), (48d)\right\}, \ f_j(\boldsymbol{x}_j) = c_j\left(\boldsymbol{p}_j^G\right),$$

and write (48c) and (48e) as (1b) and (1c), respectively.

Consider the three-area power system shown in Fig. 1 that comprises three IEEE 118-bus test systems stitched together with 6 tie-lines as shown. The three systems were modified as delineated in Appendix B.1. We applied Algorithm 1 on a reformulation of $\mathcal{P}_1$ as an instance of $\mathcal{P}$ with a flat start ($z_j(1) = 0$, $Z_j(0) = 0$, $j = 1, \ldots, N$) and step size $\eta = \eta_0/\sqrt{T}$, where $\eta_0 = 10^2$ and $T = 10^6$. The results are portrayed on the left of Fig. 2. We chose W based on the transition probabilities of a Markov chain in the Metropolis-Hastings algorithm (see [59, Sec. 2.5]). Here, $\mathcal{P}^\star$ was computed by
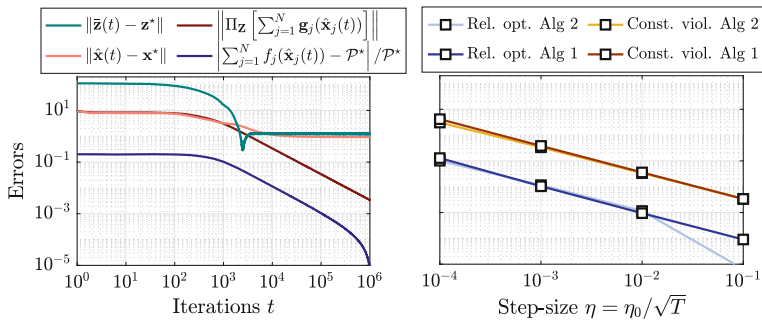
Fig. 3 Performance of Algorithm 2 with primal averaging (left) and the impact of step-size on Algorithms 1 and 2 with primal averaging (right)

solving $\mathcal{P}_1$ as a linear program. Our simulations were performed in MATLAB 2018b. All subproblems were solved via MOSEK version 9.2.17.

Algorithm 2: Distributed dual subgradient to solve $\mathcal{P}$.

---

1   Choose $z_j(1) = 0$ and $\eta = \eta_0/\sqrt{T}$.
2   **for** $t = 1, \ldots, T$ **do**
3     $x_j(t) \leftarrow \operatorname{argmin}_{x_j \in \mathbb{X}_j} \mathcal{L}_j(x_j, z_j(t))$.
4     $z_j(t+1) \leftarrow \sum_{k=1}^{N} W_{jk} \pi_{\mathbb{Z}} [z_k(t) + \eta g_k(x_k(t))]$.
5   **end**

We compared Algorithm 1 with the classical dual subgradient method in Algorithm 2 (we remark that the projection and the consensus operations in step 4 are sometimes reversed, e.g., see [63]). The progress of Algorithm 2 with the same step-size used for Algorithm 1 is shown in the right of Fig. 2. Note that Algorithm 1 leads to much smoother progress of $\sum_{j=1}^{N} f_j(x_j(t))$ compared to that with Algorithm 2. Classical dual subgradient with primal averaging via $\hat{x}_j(t) := \frac{1}{t} \sum_{r=1}^{t} x_j(r)$ for each $j = 1, \ldots, N$ can prevent this "flutter" (see [54, Sect. 4]), as the left plot in Fig. 3 reveals. While step 4 of Algorithm 1 executes a similar averaging operation, this averaging step cannot be viewed as an *output* of the iteration dynamics as is the case for Algorithm 2 with averaging. As a result, the last iterate of Algorithm 1 moves smoothly as opposed to Algorithm 2. Such an update is useful in applications that require iterates to be directly implemented as control actions and the dual subgradient is only available at the current iterate (see [57] for a detailed discussion).

In the right subgraph of Fig. 3, we compare the impact of step-size on the performance of Algorithms 1 and 2 with primal averaging after $T = 10^6$ iterations. Here, relative optimality measures $\left| \sum_{j=1}^{N} f_j(x_j(t)) - \mathcal{P}^\star \right| / \mathcal{P}^\star$ and constraint violation measures $\left\| \pi_{\mathbb{Z}}[\sum_{i=1}^{N} g_j(x_j(T))] \right\|$ for Algorithm 1 and the same at $\hat{x}_j(t)$'s for Algorithm 2. Empirically, constraint violation for Algorithm 1 appears similar to that for Algorithm 2 with primal averaging that is known to have $\mathcal{O}(T^{-1/2})$ decay rate,

much better than that suggested by (16) for Algorithm 1. We conclude this section with a remark. Notice that this algorithm takes a relatively large number of iterations to converge ($\sim 10^6$ for our example), a feature common to first-order algorithms. Such iteration count is perhaps only meaningful when the decentralized primal subproblems are easy to solve. When that is not the case, one should judiciously weigh the merits of decentralization via dual subgradient methods against faster second-order methods and a centralized solution architecture, when possible.

5 Grid Optimization Example 2: DER Coordination

Our next application problem is the coordination of DERs such as thermostatically controlled loads, electric vehicles, distributed rooftop solar, etc. that are increasingly getting adopted in distribution grids. There is a long literature on DER coordination to fulfill a variety of objectives that range from tracking a regulation signal at the T&D interface, to volt/VAR control within the distribution grid, etc. (e.g., see [14, 16]).

We formulate the DER coordination problem $\mathcal{P}_2$ over a balanced three-phase radial distribution network on N buses, described by graph $\mathfrak{G}(N, \mathbb{E})$. Let the first bus be the T&D interface. Associate directions to edges in $\mathbb{E}$ arbitrarily to obtain a directed graph $\vec{\mathfrak{G}}(N, \vec{\mathbb{E}})$, where $j \to k \in \vec{\mathbb{E}}$ denotes a directed edge from bus j to bus k in $\mathfrak{G}$. At each bus j, consider a dispatchable asset capable of injecting real and reactive powers p_j^G, q_j^G, respectively. Let $c_j(p_j^G, q_j^G)$ denote the cost of power procurement from that dispatchable asset. At $j = 1$, this cost might reflect the cost of procuring power from the transmission grid. It can also encode deviation of the power injection from a set point defined by a frequency regulation signal. At buses $j = 2, \ldots, N$, the cost can encode the disutility of deferred demand or cost of power production from dispatchable generation. The power injection capabilities of this asset at bus j are limited as $\underline{p}_j^G \le p_j^G \le \overline{p}_j^G$ along with

$$\underline{q}_j^G \le q_j^G \le \overline{q}_j^G \text{ or } \left[p_j^G \right]^2 + \left[q_j^G \right]^2 \le \left[\overline{s}_j^G \right]^2,$$

henceforth denoted as $(p_j^G, q_j^G) \in \$_j$. Such models encompass photovoltaic and energy storage systems, water pumps, commercial heating, ventilation, and air conditioning (HVAC) systems, etc. At each bus j, also assume nominal real and reactive power demands p_j^D and q_j^D.

We need additional notation to describe the DER coordination problem. Associate with bus j the squared voltage magnitude w_j, deemed to lie in $[\underline{w}_j, \overline{w}_j]$. Let $P_{j,k}$, $Q_{j,k}$ denote the real and reactive power flows from bus j to bus k for $j \to k$ in $\vec{\mathfrak{G}}$. Denote by $\ell_{j,k}$, the squared current magnitude flowing from bus j to bus k, upper bounded by $L_{j,k}$. Let $r_{j,k}$ and $\varkappa_{j,k}$ denote the resistance and reactance of the line $j \to k$. The DER coordination problem with a second-order conic relaxation of the feasible set induced by the power flow equations in the radial distribution network can be formulated as

$$\mathcal{P}_2 : \text{minimize} \quad \sum_{j=1}^{N} c_j(p_j^G, q_j^G),$$

subject to

$$\left(p_j^G, q_j^G\right) \in \mathbb{S}_j, \tag{49a}$$

$$p_j^G - p_j^D = \sum_{k:j \to k} P_{j,k} - \sum_{k:k \to j} (P_{k,j} - r_{k,j}\ell_{k,j}), \tag{49b}$$

$$q_j^G - q_j^D = \sum_{k:j \to k} Q_{j,k} - \sum_{k:k \to j} (Q_{k,j} - x_{k,j}\ell_{k,j}), \tag{49c}$$

$$w_k = w_j - 2(r_{j,k}P_{j,k} + x_{j,k}Q_{j,k}) + (r_{j,k}^2 + x_{j,k}^2)\ell_{j,k}, \tag{49d}$$

$$\ell_{j,k} \le L_{j,k}, \ \underline{w}_j \le w_j \le \overline{w}_j, \tag{49e}$$

$$\ell_{j,k}w_j \ge P_{j,k}^2 + Q_{j,k}^2,$$

$$j = 1, \ldots, N, j \to k \in \vec{\mathfrak{G}}. \tag{49f}$$

The last inequality is a second-order cone constraint, making (49) a SOCP. Constraints in (49b), (49c) describe real and reactive power balance at each bus. Relations (49b), (49c), (49c), and the inequality in (49f) replaced with an equality define the feasible set described by AC power flow equations (see [23, 49] for details).

To cast $\mathcal{P}_2$ as an instance of $\mathcal{P}$, we first write the out-neighbors of j in $\vec{\mathfrak{G}}$ as $k_1, \ldots, k_J$ and identify

$$\boldsymbol{x}_j := \left(p_j^G, q_j^G, w_j, P_{j,k_1}, \ldots, P_{j,k_J}, Q_{j,k_1}, \ldots, Q_{j,k_J}, \ell_{j,k_1}, \ldots, \ell_{j,k_J}\right)^\mathsf{T},$$

$$\mathbb{X}_j := \{\boldsymbol{x}_j \mid (49a), (49e), (49f)\}, \ f_j(\boldsymbol{x}_j) = c_j(p_j^G, q_j^G).$$

Then, it is straightforward to write (49b), (49c) and (49d) as examples of (1b). This formulation does not require inequality constraints of the form (1c). Note T&D interface's energy balancing and voltage constraints (49b)-(49d) can be treated as (1d) since there is no coupling variable from adjacent agents.

We ran Algorithm 1 on $\mathcal{P}_2$ over a modified IEEE 4-bus radial distribution network (see Appendix B.2) for details).

System conditions in the distribution grid can change quite fast. One line of work on DER coordination solves optimization problems in quick successions to deal with such changes, e.g., in [71]. To illustrate the use of DER coordination with time-varying distribution grid conditions, we simulated a case where real and reactive power demands were changed every 10^6 iterations as prescribed in (Appendix B.2) with step-size $\eta = 0.1$. Algorithm 1 is restarted after every change. Here, we use the last primal-dual iterate at the point of change to restart Algorithm 1. As Fig. 4b illustrates, Algorithm 1 can track the optimal cost in the changing problem environment.

Convergence slows down over larger networks. Consider a modified IEEE 15-bus radial distribution network; see Appendix B.2 for the modifications. Figure 6 illustrates that Algorithm 1 with restarts is able to track optimal costs, but only when nodes are
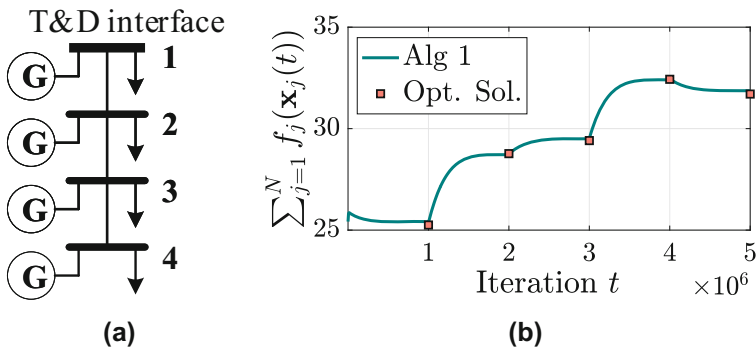
Fig. 4 **a** A 4-bus radial network. **b** Progress of the objective function at the last iterate of Algorithm 1 on $\mathcal{P}_2$ for the network in Fig. 4a
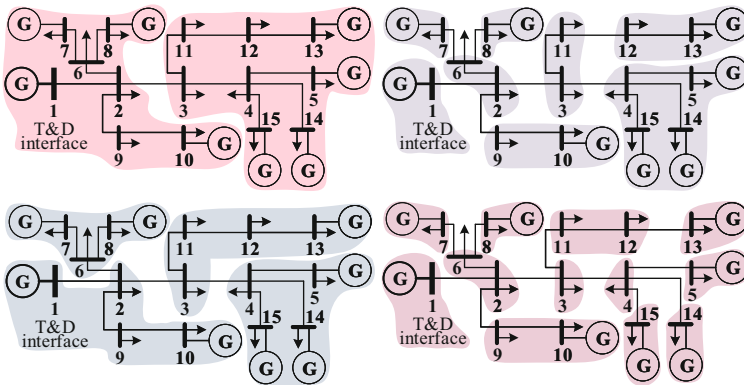
Fig. 5 The IEEE 15-bus test feeder subdivided into 2, 4, 8 and 12 groups
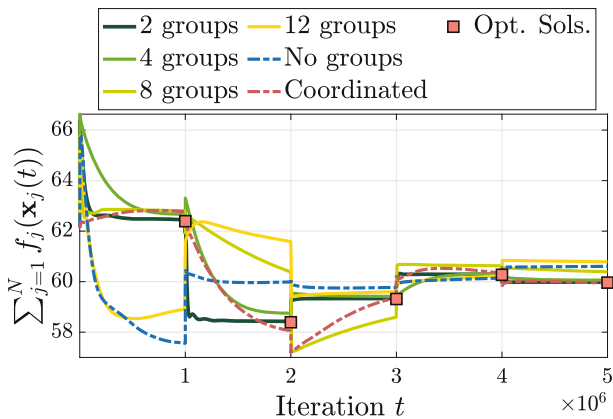
Fig. 6 Evolution of the objective function value of Algorithm 1 on the IEEE 15-bus test system with varying degrees of decentralization (based on groupings of buses per Fig. 5). Performance of the coordinated DDSG method with averaging in [48, Algorithm 1] with constant step-size is shown in the red dashed line

grouped together into fewer agents in $\mathcal{P}$, as shown in Fig. 5. Such a slowdown is expected, given that the convergence guarantees depend on N, albeit polynomially. We remark that *all* first-order algorithms suffer from this issue. In practical implementation, one must carefully explore the trade-off between the degree of decentralization and the accuracy of tracking within a fixed number of iterations. To see the merits of a more centralized architecture, consider the algorithm in [48, Algorithm 1] with constant step-size that corresponds to the same algorithm as DDSG, but where a central coordinator updates the dual variables for all primal subproblems simultaneously. It tracks the optimal solution faster than all possible groupings of the buses, per Fig. 6. Thus, coordination improves speed of convergence. Our study highlights that centralization of solution architecture, possibly through coordination or otherwise, is often more suited for real-time tracking problems with stringent speed requirements.

We conclude this section with two remarks. First, notice that the solution architecture we study stipulates each primal subproblem to be solved at each iteration. This subproblem contributes to the knowledge of a local subgradient of the dual function. Despite needing more coordination, bundle methods proposed in [65] have the advantage that not all subproblems need to be solved at each iteration and this flexibility is appealing when the number of agents is large and the computational cost for solving the primal subproblems can be high. These methods utilize the whole history of evaluations of the dual function and the subdifferential set rather than an arbitrary local subgradient to update the current iterate; see [15] for details. A meaningful comparison and possible merger between bundle methods and our algorithms with applications to DER coordination is left for future work. Second, notice that our formulation in (49) considers a deterministic optimization problem over a single snapshot in time. Our solution architecture should apply to the multi-stage counterpart with stochastic supply and energy storage as considered in [27], whose formulation with SOCP relaxation parallels ours and comes with tightness guarantees for said relaxation.

6 Grid Optimization Example 3: T&D Coordination

The lack of visibility of transmission system operators (SOs) into distribution grids and bottlenecks in wholesale market clearing software make it impossible for such SOs to directly harness the flexibility offered by DERs in the distribution networks. Naturally, distributed algorithms are suited for T&D coordination. Assume that for each distribution network, an aggregator $\mathcal{A}$ directly controls the dispatchable DERs and knows the network parameters. In what follows, we present the T&D coordination problem $\mathcal{P}_3$ that a transmission SO and a collection of distribution aggregators solve in a distributed fashion. We utilize a semidefinite relaxation of the feasible set induced by the power flow equations for the transmission network and a linear distribution flow model for the distribution grids.

To formulate the joint dispatch problem of all T&D assets, we require three different graphs. The first among these is the transmission network, modeled as an undirected graph $\mathfrak{G}^{\text{tran}}$ on n^{tran} transmission buses. The second set of graphs is for the distribution grids that connect to the transmission network at their points of common coupling–the n^{tran} transmission buses. We model the distribution grid connected to transmission bus

ℓ as an undirected graph $\mathfrak{G}_\ell^{\mathrm{dist}}$ on $n_\ell^{\mathrm{dist}} + 1$ distribution buses, where the first bus of $\mathfrak{G}_\ell^{\mathrm{dist}}$ coincides with bus ℓ in $\mathfrak{G}^{\mathrm{tran}}$. Finally, we consider an undirected star graph $\mathfrak{G}$ on $N = n^{\mathrm{tran}} + 1$ nodes with the aggregators $\mathcal{A}_1, \ldots, \mathcal{A}_{n^{\mathrm{tran}}}$ as the satellite nodes and the SO (the N-th node) at the center.

Let $V \in \mathbb{C}^n$ denote the vector of nodal voltage phasors, where $\mathbb{C}$ is the set of complex numbers. We formulate the engineering constraints of the grid using the positive semidefinite matrix $\Lambda := VV^{\mathsf{H}} \in \mathbb{C}^{n^{\mathrm{tran}} \times n^{\mathrm{tran}}}$. To describe these constraints, let $y_{\ell,k} = y_{k,\ell}$ denote the admittance of the transmission line joining buses ℓ, k in $\mathfrak{G}^{\mathrm{tran}}$ and $y_{\ell,\ell}$ denote the shunt admittance at bus ℓ. Then, define $\Phi_{\ell,k}, \Psi_{\ell,k}$ as the $n^{\mathrm{tran}} \times n^{\mathrm{tran}}$ Hermitian matrices whose only nonzero entries are

$$[\Phi_{\ell,k}]_{\ell,\ell} := \frac{1}{2}(y_{\ell,k} + y_{\ell,k}^{\mathsf{H}}), \; [\Phi_{\ell,k}]_{\ell,k} = [\Phi_{\ell,k}]_{k,\ell}^{\mathsf{H}} := -\frac{1}{2}y_{\ell,k},$$

$$[\Psi_{\ell,k}]_{\ell,\ell} := \frac{1}{2i}(y_{\ell,k}^{\mathsf{H}} - y_{\ell,k}), \; [\Psi_{\ell,k}]_{\ell,k} = [\Psi_{\ell,k}]_{k,\ell}^{\mathsf{H}} := \frac{1}{2i}y_{\ell,k}.$$

In addition, we define the $n^{\mathrm{tran}} \times n^{\mathrm{tran}}$ Hermitian matrices

$$\Phi_\ell := \frac{1}{2}\left(y_{\ell,\ell} + y_{\ell,\ell}^{\mathsf{H}}\right)\mathbb{1}_\ell\mathbb{1}_\ell^{\mathsf{H}} + \sum_{k \sim \ell}\Phi_{\ell,k}, \quad \Psi_\ell := \frac{1}{2i}\left(y_{\ell,\ell}^{\mathsf{H}} - y_{\ell,\ell}\right)\mathbb{1}_\ell\mathbb{1}_\ell^{\mathsf{H}} + \sum_{k \sim \ell}\Psi_{\ell,k},$$

where $\mathbb{1}$ is a vector of all ones of appropriate size and $\mathbb{1}_\ell$ is a vector of all zeros except at the ℓ-th position that is unity. This notation allows us to describe the apparent power flow from bus ℓ to bus k as $\mathrm{Tr}(\Phi_{\ell,k}\Lambda) + i\,\mathrm{Tr}(\Psi_{\ell,k}\Lambda)$, the apparent power injection at bus ℓ as $\mathrm{Tr}(\Phi_\ell\Lambda) + i\,\mathrm{Tr}(\Psi_\ell\Lambda)$, and the squared voltage magnitude at bus ℓ as $\mathrm{Tr}(\mathbb{1}_\ell\mathbb{1}_\ell^{\mathsf{H}}\Lambda)$, where $i := \sqrt{-1}$. At each transmission bus ℓ with load $P_\ell^D + iQ_\ell^D$, let a generator supply apparent power $P_\ell^G + iQ_\ell^G$ with procurement cost described by C_ℓ.

Let $p_\ell + iq_\ell \in \mathbb{C}^{n^{\mathrm{dist}}}$ denote the vector of net power injections across the distribution network, save the first bus. Further, let the power procurement cost be given by c_ℓ to inject $p_\ell + iq_\ell \in \mathbb{C}^{n^{\mathrm{dist}}}$. Also, let $\Lambda_\ell \in \mathbb{C}^{n^{\mathrm{dist}}}$ denote the vector of squared voltage magnitudes across the same set of buses. We adopt the popular LinDistFlow model to tackle the nonconvex nature of the power flow equations in the distribution grid. Let $\widetilde{M} \in \mathbb{R}^{n^{\mathrm{dist}} \times n^{\mathrm{dist}}}$ be the node-to-edge incidence matrix of $\mathfrak{G}_\ell^{\mathrm{dist}}$. Further, remove the first row of $\widetilde{M}$ to obtain the reduced incidence matrix M. Then, the voltage magnitudes are related to power injections under the LinDistFlow model as $\Lambda_\ell = \rho_\ell p_\ell + \chi_\ell q_\ell + \Lambda_{\ell,\ell}\mathbb{1}$, where ρ_ℓ and χ_ℓ are $n^{\mathrm{dist}} \times n^{\mathrm{dist}}$ matrices defined as $\rho_\ell := 2M^{-\mathsf{T}}\mathrm{diag}(r_\ell)M^{-1}$, $\chi_\ell := 2M^{-\mathsf{T}}\mathrm{diag}(\chi_\ell)M^{-1}$, r_ℓ/χ_ℓ collect the resistances/reactances of the n^{dist} distribution lines.

The optimal joint dispatch of all T&D assets is given by

$$\mathcal{P}_3 : \text{minimize} \qquad \sum_{\ell=1}^{n^{\mathrm{tran}}} C_\ell(P_\ell^G, Q_\ell^G) + \sum_{\ell=1}^{n^{\mathrm{tran}}} c_\ell(p_\ell^G, q_\ell^G),$$

$$\text{subject to} \qquad (P_\ell^G, Q_\ell^G) \in \mathbb{S}_\ell^{\mathrm{tran}}, \tag{50a}$$

$$\left(\boldsymbol{p}_\ell^G, \boldsymbol{q}_\ell^G \right) \in \mathbb{S}_\ell^{\text{dist}}, \tag{50b}$$

$$P_\ell^G - P_\ell^D + \mathbb{1}^{\mathsf{T}} \left(\boldsymbol{p}_\ell^G - \boldsymbol{p}_\ell^D \right) = \text{Tr}(\boldsymbol{\Phi}_\ell \boldsymbol{\Lambda}), \tag{50c}$$

$$Q_\ell^G - Q_\ell^D + \mathbb{1}^{\mathsf{T}} \left(\boldsymbol{q}_\ell^G - \boldsymbol{q}_\ell^D \right) = \text{Tr}(\boldsymbol{\Psi}_\ell \boldsymbol{\Lambda}), \tag{50d}$$

$$\text{Tr}(\boldsymbol{\Phi}_{\ell,\ell'} \boldsymbol{\Lambda}) \leq L_{\ell,\ell'}, \tag{50e}$$

$$\underline{\Lambda}_{\ell,\ell} \leq \Lambda_{\ell,\ell} \leq \overline{\Lambda}_{\ell,\ell}, \tag{50f}$$

$$\boldsymbol{\Lambda} \succeq 0, \tag{50g}$$

$$\underline{\boldsymbol{w}}_\ell \leq \boldsymbol{\rho}_\ell \boldsymbol{p}_\ell + \boldsymbol{\chi}_\ell \boldsymbol{q}_\ell + \Lambda_{\ell,\ell} \mathbb{1} \leq \overline{\boldsymbol{w}}_\ell, \tag{50h}$$

$$\text{for } \ell = 1, \ldots, n^{\text{tran}}, \ \ell' \sim \ell.$$

Here, (50a), (50b) encode the capabilities of the transmission and the distribution assets, respectively, where $\mathbb{S}^{\text{tran}}$ and $\mathbb{S}^{\text{dist}}$ are assumed convex. Constraints in (50c) and (50d) enforce nodal power balance at transmission nodes. Transmission line (real power) flows are constrained in (50e), with L's describing the line capacities. One can alternately constrain apparent power flows. Transmission voltage limits are enforced via (50f) within $[\underline{\Lambda}_{\ell,\ell}, \overline{\Lambda}_{\ell,\ell}]$. The relation $\boldsymbol{\Lambda} = \boldsymbol{V} \boldsymbol{V}^{\mathsf{H}}$ requires $\boldsymbol{\Lambda}$ to be positive semidefinite (enforced in (50g)) and rank-1. We consider the rank-relaxed power flow equations by dropping the rank-1 requirement. For the distribution grid, voltage limits are enforced in (50h) and power flow limits are ignored.

Recall that $\mathfrak{G}$ for T&D coordination problem is a graph on $N = n^{\text{tran}} + 1$ nodes, where the first n^{tran} nodes are transmission buses and the last node represents the SO. Define

$$\boldsymbol{x}_\ell := \left(\boldsymbol{p}_\ell^{\mathsf{T}}, \boldsymbol{q}_\ell^{\mathsf{T}} \right)^{\mathsf{T}}, \ \mathbb{X}_\ell := \{ \boldsymbol{x}_\ell \mid (50b) \}, \ f_\ell = c_\ell(\boldsymbol{p}_\ell^G, \boldsymbol{q}_\ell^G)$$

for $\ell = 1, \ldots, n^{\text{tran}}$. Collect the real and reactive power generations across the transmission grid in the vectors $\boldsymbol{P}^G$, $\boldsymbol{Q}^G$, respectively. Then, define

$$\boldsymbol{x}_N := \left([\boldsymbol{P}^G]^{\mathsf{T}}, [\boldsymbol{Q}^G]^{\mathsf{T}}, \text{vec} \left(\Re\{\boldsymbol{\Lambda}\}^{\mathsf{T}} \right), \text{vec} \left(\Im\{\boldsymbol{\Lambda}\}^{\mathsf{T}} \right) \right)^{\mathsf{T}},$$

$$\mathbb{X}_N := \{ \boldsymbol{x}_{n^{\text{tran}}+1} \mid (50a), (50e), (50f), (50g) \},$$

$$f_N(\boldsymbol{x}_N) = \sum_{\ell=1}^{n^{\text{tran}}} C_\ell(P_\ell^G, Q_\ell^G).$$

The constraint (50g) can be written in terms of $\boldsymbol{x}_N$ as

$$\begin{pmatrix} \Re\{\boldsymbol{\Lambda}\} & \Im\{\boldsymbol{\Lambda}\} \\ -\Im\{\boldsymbol{\Lambda}\} & \Re\{\boldsymbol{\Lambda}\} \end{pmatrix} \succeq 0$$

and (50c) – (50d) as examples of (1b) using $\text{Tr}(\boldsymbol{\varphi} \boldsymbol{\Lambda}) = \text{vec} (\Re\{\boldsymbol{\varphi}\})^{\mathsf{T}} \text{vec} (\Re\{\boldsymbol{\Lambda}\}) + \text{vec} (\Im\{\boldsymbol{\varphi}\})^{\mathsf{T}} \text{vec} (\Im\{\boldsymbol{\Lambda}\})$ for a Hermitian matrix $\boldsymbol{\varphi}$. Constraints (50h) are examples of inequality constraints in (1c).
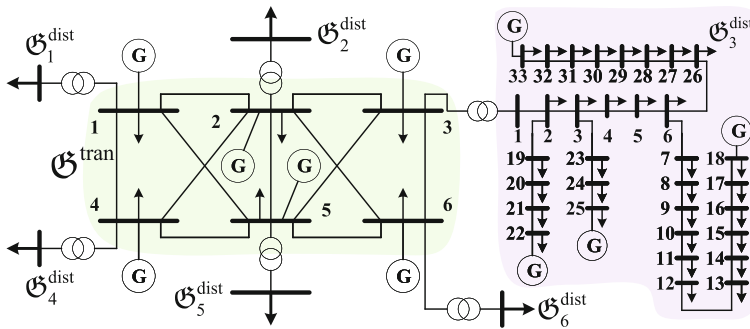
Fig. 7 The 204-bus network for T&D simulations, obtained by joining the IEEE 6-bus transmission network with six IEEE 33-bus distribution networks
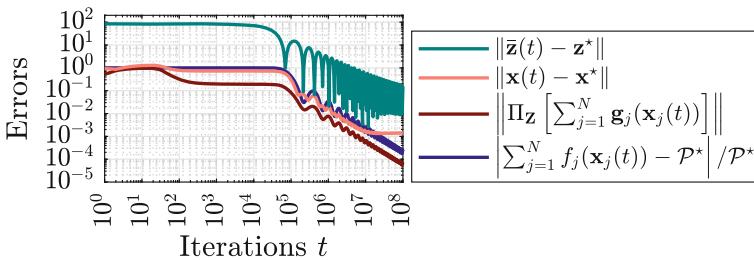
Fig. 8 Progress of Algorithm 1 on $\mathcal{P}_3$

We report numerical results on a 204-bus T&D system that comprises the IEEE 6-bus transmission network joined with six IEEE 33-bus distribution systems (see Fig. 7 and Appendix B.3 for details). We applied Algorithm 1 on a reformulation of $\mathcal{P}_3$ as an instance of $\mathcal{P}$ with a flat start ($z_j(1) = 0$, $\mathbf{Z}_j(0) = 0$, $j = 1, \ldots, N$) and step size $\eta = \eta_0/\sqrt{T}$, where $\eta_0 = 10^3$ and $T = 10^8$. The agent-wise subproblems for $\mathcal{P}_3$ are communicated over a 7-node star graph $\mathfrak{G}$ with the SO in the center. Convergence results are shown in Fig. 8.

One might surmise that dual subgradient methods can be accelerated à la Nesterov. Such acceleration requires smoothness of the dual function–a property that an application problem such as $\mathcal{P}_3$ does not always possess. In Appendix A, we provide simple examples, where dual functions are nonsmooth, and show how the nonsmoothness impedes acceleration. We illustrate the same difficulty with $\mathcal{P}_3$. Specifically, we compare the performance of Algorithm 1 with a distributed acceleration scheme in [28, Eqs. (5)-(6)], adopted to our setup, described in Algorithm 3.

In this algorithm, the dual ancillary sequences $\mathbf{Z}_j(t)$, $\mathbf{Y}_j(t)$ in steps 3-4 are averaged by $\alpha(t)$ in step 6, where $\alpha(t)$ is updated in step 5. The distributed acceleration scheme in [28, Eqs. (5)-(6)] has been designed for unconstrained optimization problems; we project multipliers of constraints (50h) in step 6 on the positive orthant to apply it to $\mathcal{P}_3$. Each local problem is solved in step 7 to obtain the local dual subgradient $\mathbf{g}_j(\mathbf{x}_j(t+1))$. The central subgradient is approximated in step 8 that combines consensus tracking and local gradient averaging.

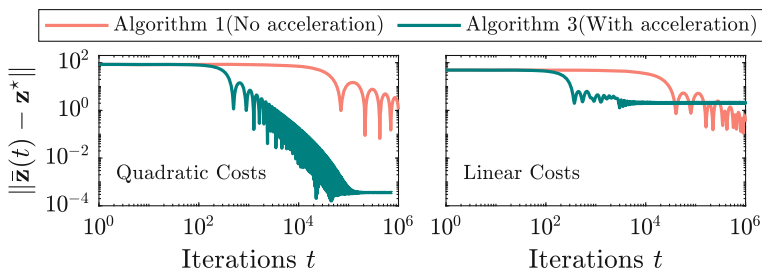**Algorithm 3:** Accelerated distributed dual subgradient to solve $\mathcal{P}$.

1 Choose $\eta = \eta_0/\sqrt{T}$, $z_j(1) = Y_j(1) = 0$, $\alpha(1) = 0.5$, $x_j(1) \leftarrow \mathrm{argmin}_{x_j \in \mathbb{X}_j} \mathcal{L}_j(x_j, z_j(1))$, and
   $s_j(1) \leftarrow g_j(x_j(1))$.

2 **for** $t = 1, \ldots, T$ **do**

3    $\quad Z_j(t+1) \leftarrow \sum_{k=1}^N W_{jk} z_k(t) + \eta s_j(t)$

4    $\quad Y_j(t+1) \leftarrow \sum_{k=1}^N W_{jk} Y_k(t) + \frac{\eta}{\alpha(t)} s_j(t)$

5    $\quad \alpha(t+1) \leftarrow \frac{1}{2}\left[-\alpha(t)^2 + \sqrt{\alpha(t)^4 + 4\alpha(t)^2}\right]$

6    $\quad z_j(t+1) \leftarrow \pi_{\mathbb{Z}}\left[(1 - \alpha(t+1))Z_j(t+1) + \alpha(t+1)Y_j(t+1)\right]$

7    $\quad x_j(t+1) \leftarrow \mathrm{argmin}_{x_j \in \mathbb{X}_j} \mathcal{L}_j(x_j, z_j(t+1))$

8    $\quad s_j(t+1) \leftarrow \sum_{k=1}^N W_{jk} s_k(t) + g_j(x_j(t+1)) - g_j(x_j(t))$

9 **end**

Fig. 9 Convergence of dual iterates on $\mathcal{P}_3$ about Algorithm 1 (no acceleration) and Algorithm 3 from [28, Eqs. (5)-(6)] (with acceleration)

The results are shown in Fig. 9. Here, we adopted cold start for the dual iterates for both algorithms, i.e., $z_j(1) = 0$, $j = 1, \ldots, N$. We show the case when generation costs for real power are considered quadratic on the left. The figure on the right is derived with linear real power generation costs. All reactive generation costs are considered quadratic for both simulations. Quadratic costs often yield a smooth dual function that Algorithm 3 can exploit and attain faster convergence than Algorithm 1. With linear costs, the dual function can be nonsmooth (especially at an optimal dual solution) as a result of primal non-uniqueness [6, Theorem 6.3.3]. Nonsmoothness is a fundamental barrier to acceleration as our example reveals. We discuss the role of smoothness of the dual function in acceleration further in Appendix A through illustrative examples.

7 Selective Prior Works on Distributed Grid Optimization

The literature on distributed optimization techniques applied to grid optimization problems $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$ is extensive and growing. While our goal in Sects. 4-6 was to illustrate the properties of the DDSG algorithm we considered, we include a partial survey in Table 1 for the interested reader to follow up on.

Perhaps the most widely studied algorithms are those based on dual decomposition. Very similar to our approach, agents are often required to agree on optimization vari-

Table 1 Some papers on distributed algorithms for $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$

Method	Problem and citations
Dual subgradient	$\mathcal{P}_1$ [2, 3, 19, 21, 29, 41, 44, 46], $\mathcal{P}_2$ [60, 66]
Saddle point dynamics	$\mathcal{P}_1$ [1, 32, 67], $\mathcal{P}_2$ [16]
Primal gradient	$\mathcal{P}_1$ [4, 40], $\mathcal{P}_2$ [8, 13, 18, 26]
Cutting plane, column and constraint generation	$\mathcal{P}_1$ [45, 70], $\mathcal{P}_3$ [69]
Multi-parametric programming	$\mathcal{P}_1$ [30, 35, 36]

ables that correspond to power flows, voltage phase angles, etc. Constraints enforcing them are dualized and the primal/dual variables are then updated suitably. Techniques such as quadratic regularization of the dual function, inexact dual (sub)gradient updates, and heuristic adjustments to these updates have been explored. These works include both dual (sub)gradient and primal-dual saddle-point dynamics. A notably distinct class of algorithms for power system problems include those where primal variables common to multiple agents are updated by a coordinator in an outer loop, while a primal subproblem in the inner-loop optimizes other variables with the common ones fixed. The art in these algorithms lies in exploiting the nature of the constraints and the information obtained from the primal subproblems to update the common variables, either using Benders decomposition, column or constraint generation techniques, or using multi-parametric programming.

8 Concluding Remarks

We have studied a constant step-size distributed dual subgradient (DDSG) method with averaging that provides order-optimal $1/\sqrt{T}$ convergence rate after T iterations for multi-agent convex-constrained optimization problems of the form $\mathcal{P}$ whose objective function need not be strongly convex. The convergence guarantee improves the rate of the decaying step-size, fully decentralized counterpart studied in [48]. We formulated a variety of problems that arise in the operation of an electric power system as examples of $\mathcal{P}$ and applied the DDSG method to solve them. The case studies with DDSG on grid optimization problems are presented not to empirically challenge other methods in the literature, but as a means to illustrate the broad applicability of optimization problems such as $\mathcal{P}$ to facilitate unified algorithm development. These examples also underscore various properties of the DDSG algorithm. Specifically, the first example illustrates the difference in performance of the last iterate between the vanilla DDSG method and ours. The second example reveals the effect of the number of agents on convergence speed. The third example empirically studies the possibility of Nesterov-style acceleration in DDSG methods and the role that smoothness in the dual function plays in such analysis.

Several research directions are of interest to us. The first among them is an algorithm with last-iterate guarantees over time-varying communication networks. The second important direction is a disentanglement of the error $\mathscr{V}_T$ to provide guarantees on sub-

optimality and constraint violation separately. A third direction is the development of similar algorithmic architectures for stochastic and risk-sensitive convex optimization problems.

A Appendix A: Dual Subgradient Methods Cannot Generally be Accelerated

Nesterov-type acceleration relies on smoothness of the objective function. Here, we illustrate why such acceleration is generally untenable in dual subgradient settings. We first present examples where the dual function for constrained optimization problems are nonsmooth. Then, we illustrate through examples how nonsmoothness of the objective function impairs acceleration. Consider an optimization problem of the form

$$\underset{x \in \mathbb{X}}{\text{minimize}} \ \frac{1}{2} x^{\mathsf{T}} \Xi x + \xi^{\mathsf{T}} x + \nu, \ \ \text{subject to } Ax \le b. \tag{51}$$

with Ξ being positive semidefinite (but not positive definite) and $\mathbb{X}$ being a convex polyhedral set. This is a quadratic program (QP) that simplifies to a linear program (LP) when $\Xi = 0$. Associate multipliers $z \ge 0$ with $Ax \le b$. Then, the dual function for the problem in (51) is given by

$$\mathcal{D}(z) := -b^{\mathsf{T}} z + \nu + \underset{x \in \mathbb{X}}{\text{minimum}} \ \left\{ \frac{1}{2} x^{\mathsf{T}} \Xi x + \xi^{\mathsf{T}} x + z^{\mathsf{T}} Ax \right\}. \tag{52}$$

With parameter choices

$$
\mathbb{X} := [0, 0.1]^3, \ \xi = - \begin{bmatrix} 17 \ 17 \ 11 \end{bmatrix}^{\mathsf{T}}, \ \nu = 5,
$$
$$
A = \begin{bmatrix} 0.19 \ 0.12 \ 0.42 \\ 0.37 \ 0.54 \ 0.13 \end{bmatrix}, \ b = \begin{bmatrix} 0.04 \\ 0.06 \end{bmatrix} \tag{53}
$$

we plot $\mathcal{D}(z)$ with $\Xi = 0$ and $\Xi = \text{diag} \, (24, 26, 0)$, respectively, in the left and right of Fig. 10. The primal and dual optimum is $\mathcal{P}^\star = \mathcal{P}_D^\star = 2.43$ for the QP and 2.30 for the LP.

$\mathcal{D}$ in Fig. 10 is nonsmooth. One might expect that for QP with $\Xi \ne 0$, $\mathcal{D}$ might be smooth. Indeed with positive definite Ξ, $\mathcal{D}$ becomes smooth, per [53, Lemma 2.2] or [6, Theorem 6.3.3]. Smoothness can no longer be guaranteed, when Ξ is only positive semidefinite, as evidenced by Fig. 10.
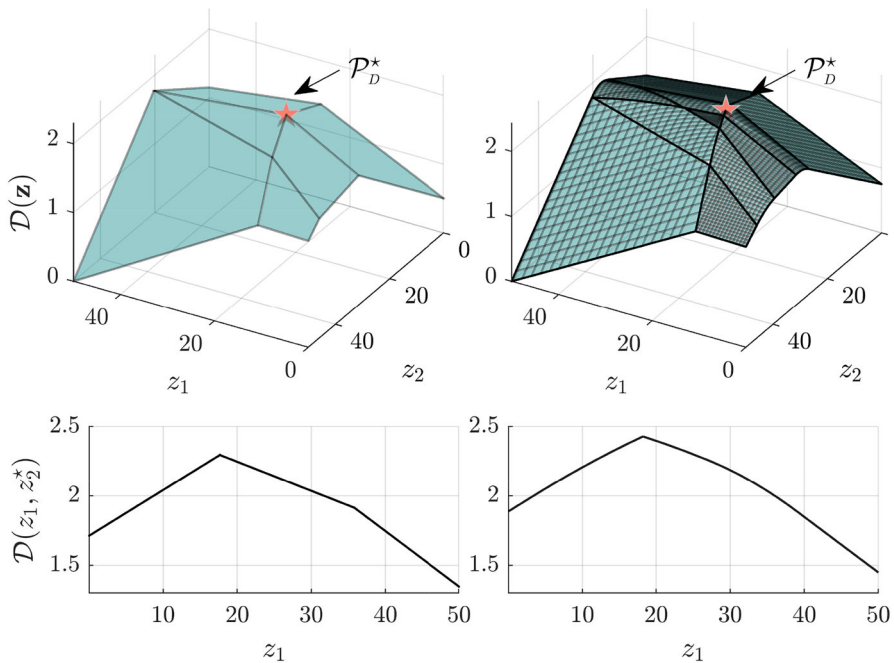
Fig. 10 Illustrations of nonsmooth $\mathcal{D}(z)$ on top and a slice $\mathcal{D}(\cdot, z_2^\star)$ on the bottom for (51) with parameters in (53). The plots on the left are obtained with $\Xi = 0$ and on the right with $\Xi = \mathrm{diag}\,(24, 26, 0)$

To demonstrate the impact of nonsmoothness on acceleration, consider the problem

$$\underset{x \in \mathbb{R}^2}{\text{minimize}} \quad f(x) := \frac{1}{2}\|x - x_C\|_2^2 + \lambda\|x\|_1. \tag{54}$$

With $x_C = 0$, f is minimized at the origin. If $\lambda = 0$, then f is smooth. For $\lambda > 0$, f is nonsmooth at the origin. With $x_C \neq 0$, f is again smooth with $\lambda = 0$. However, with $\lambda > 0$, f is nonsmooth at the origin, but not at the optimum of f.

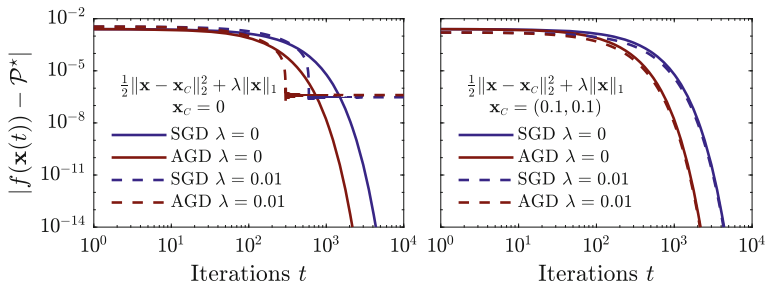We compare (sub)gradient descent, i.e., $x(t + 1) = x(t) - \eta\nabla f(x(t))$, with an accelerated variant described in Algorithm 4 on (54).

Algorithm 4: Accelerated Gradient Descent adopted from [71], Sect. 3.7.2

---

1 Choose $\eta = 0.003$, $x(1) = y(1) = (0.05, 0.05)$, and $\alpha(1) = 0$.

2 **for** $t = 1, \ldots, T$ **do**

3      $\alpha(t + 1) \leftarrow \frac{1 + \sqrt{1 + 4\alpha(t)^2}}{2}$, $\gamma(t) \leftarrow \frac{1 - \alpha(t)}{\alpha(t+1)}$

4      $y(t + 1) \leftarrow x(t) - \eta\nabla f(x(t))$

5      $x(t + 1) \leftarrow (1 - \gamma(t))y(t + 1) + \gamma(t)y(t)$

6 **end**

Fig. 11 Comparison of SGD and AGD on (54) with $x_C = 0$ on the left and with $x_C = (0.1, 0.1)$ on the right

We start both algorithms at $(0.05, 0.05)$ with $\eta = 0.003$. For accelerated (sub)gradient descent (AGD), we also set $\alpha(1) = 0$. The plot on the left side of Fig. 11 shows the progress of subgradient descent (SGD) and AGD on (54) with $x_C = 0$. Note that when $\lambda = 0$, i.e., when f is smooth, AGD outperforms SGD. However, when $\lambda = 0.01$ and f is nonsmooth at the optimum, AGD performs better initially, but both algorithms oscillate around the optimum with similar errors, eventually.

In the right subgraph of Fig. 11, we compare SGD and AGD under the same settings, but with $x_C = (0.1, 0.1)$. Note that AGD now performs better than SGD with zero and nonzero λ. That is, when the nonsmoothness is away from the optimum, AGD can accelerate convergence to the optimum locally, as long as the iterates remain within a region around the optimum where the function is locally smooth.

B Appendix B: Simulation Details for Sections 4, 5 and 6

Network data were obtained from MATPOWER 7.1.

B.1 Appendix B1: Data for Solving $\mathcal{P}_1$

The multi-area power system considered in Section 4 is illustrated in Figure 1. The 118-bus networks were modified as follows. Tie-line capacities were set to 100MW and their reactances were set to 0.25p.u. Capacities of transmission lines internal to each area were set to 100MW. All loads and generators at boundary buses were removed. Quadratic cost coefficients were neglected and the linear cost coefficients c_j of the generators were perturbed to $\widetilde{c}_j := c_j \circ (0.99 + 0.02\xi_j)$, for $j = 1, \ldots, N$, where entries of ξ_j are independent $\mathcal{N}(0, 1)$ (standard normal) variables. All phase angles were restricted to $[-\frac{\pi}{6}, \frac{\pi}{6}]$.

B.2 Appendix B2: Data for Solving $\mathcal{P}_2$

The 4-bus network considered in Section 5, shown in Figure 6, is modified from the IEEE 4-bus network as follows. The branch joining buses 1 and 4 was altered to connect buses 3 and 4. We enforced squared current flows as $\ell_{j,k} \in [0, 200]$ Amp^2,

Table 2 Cost coefficients of the 4-bus network for $\mathcal{P}_2$.

Bus	1	2	3	4
α_p[\$/MW^2]	0	6	7	8
β_p[\$/MW]	30	19	18	17
α_q[\$/MVAR^2]	5	5.1	5.2	5.3

Table 3 Cost coefficients of the 15-bus network for $\mathcal{P}_2$.

Bus	1	5	7	8	10	13	14	15
α_p[\$/MW^2]	0	25	23	21	19	17	15	13
β_p[\$/MW]	50	41	42	43	44	45	46	47
α_q[\$/MVAR^2]	25	24	23	22	21	20	19	18

Table 4 Cost coefficients of the 6-bus network for $\mathcal{P}_3$.

Bus	1	2	3	4	5	6
α_p[\$/MW^2]	8.7	5.9	6.8	7.2	4.2	3.5
β_p[\$/MW]	11	12	13	14	15	16
α_q[\$/MVar^2]	3.2	3.5	2.3	1.8	1.5	1.7

real and reactive branch power flows as $P_{j,k} \in [-1, 1]$ MW and $Q_{j,k} \in [-1, 1]$ MVAR, respectively. DER generators were added at buses 2, 3 and 4. Bus 1 defined the T&D interface. Generation capacities were fixed to $[0, 1]$ MW and $[-1, 1]$ MVAR. Generation costs were $\alpha_{pj}(p_j^G)^2 + \beta_{pj} p_j^G + \alpha_{qj}(q_j^G)^2$ with coefficients in Table 2.

For the IEEE 15-bus system shown in Figure 5, we modified the branch flow limits to mirror those for the 4-bus system. We added 7 distributed generators at buses 5, 7, 8, 10, 13, 14, 15, where bus 1 is the T&D interface, all with capacities $[0, 0.2]$ MW and $[-0.2, 0.2]$ MVAR. Generation costs were similar to the 4-bus network with coefficients in Table 3.

We randomized the real and reactive power demands at each change point by scaling each (real/reactive) load by $[\omega + (\omega' - \omega)\xi]$, where $\xi \sim \mathcal{N}(0, 1)$. Parameters (ω, ω') were varied at the change points in the sequence (0.70, 1.30), (0.80, 1.20), (0.85, 1.15), (0.75, 1.20), (0.95, 1.05). The experiment was initialized with default loads from MATPOWER.

B.3 Appendix B3: Data for Solving $\mathcal{P}_3$

In Section 6, for the 204-bus system in Figure 7, the 6-bus transmission network was modified as follows. All branch capacities are set to 5MW. All real and reactive generation capacities were set to $[0, 5]$MW and $[-5, 5]$ MVAR, respectively. We considered $P_\ell^D + jQ_\ell^D = (4 + j4)$[MVA] at each bus $\ell = 1, \ldots, 6$. Generation costs were similar to the 4-bus network with coefficients in Table 4.

For all 33-bus distribution networks, all branch capacities were set to 4 MW. Four DER generators were added at buses 18, 22, 25 and 33. Bus 1 is the T&D interface. Again, we considered generation costs as for $\mathcal{P}_2$ but with coefficients $\boldsymbol{\alpha}_{p\ell} = 5 \circ (0.9 + 0.1\boldsymbol{\xi}_\ell)$, $\boldsymbol{\beta}_{p\ell} = 20 \circ (0.9 + 0.1\boldsymbol{\xi}'_\ell)$ and $\boldsymbol{\alpha}_{q\ell} = 3 \circ (0.9 + 0.1\boldsymbol{\xi}''_\ell)$ for $\ell = 1, \ldots, n^{\text{tran}}$, where all entries of $\boldsymbol{\xi}_\ell, \boldsymbol{\xi}'_\ell, \boldsymbol{\xi}''_\ell$ are drawn from $\mathcal{N}(0, 1)$. Real and reactive power demands in the distribution networks were randomized similarly to that for $\mathcal{P}_2$ with $\omega = 0.9$ and $\omega' = 1.1$.

References

1. Ahmadi-Khatir, A., Conejo, A.J., Cherkaoui, R.: Multi-area energy and reserve dispatch under wind uncertainty and equipment failures. IEEE Trans. Power Syst. **28**(4), 4373–4383 (2013)
2. Arnold, M., Knopfli, S., Andersson, G.: Improvement of OPF decomposition methods applied to multi-area power systems. In: 2007 IEEE Lausanne Power Tech, pp. 1308–1313. IEEE, Lausanne, Switzerland (2007)
3. Bakirtzis, A.G., Biskas, P.N.: A decentralized solution to the DC-OPF of interconnected power systems. IEEE Trans. Power Syst. **18**(3), 1007–1013 (2003)
4. Baldick, R., Chatterjee, D.: Coordinated dispatch of regional transmission organizations: theory and example. Comput. Oper. Res. **41**, 319–332 (2014)
5. Baran, M.E., Felix, F.W.: Optimal capacitor placement on radial distribution systems. IEEE Trans. Power Deliv. **4**(1), 725–734 (1989)
6. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms, 3rd edn. Wiley, New York (2006)
7. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. Athena Scientific, Belmont, MA (1997)
8. Bhattarai, B.P., de Cerio Mendaza, I.D., Myers, K.S., Bak-Jensen, B., Paudyal, S.: Optimum aggregation and control of spatially distributed flexible resources in smart grid. IEEE Trans. Smart Grid **9**(5), 5311–5322 (2018)
9. Bose, S., Gayme, D.F., Mani Chandy, K., Steven, H.: Low quadratically constrained quadratic programs on acyclic graphs with application to power flow. IEEE Trans. Control Netw. Syst. **2**(3), 278–287 (2015)
10. Boyd, S., Xiao, L., Mutapcic, A., Mattingley, J.: Notes on Decomposition Methods EE364B. Stanford University, Stanford (2015)
11. Bubeck, S.: Convex optimization: algorithms and complexity. Found. Trends Mach. Learn. **8**(3–4), 231–357 (2015)
12. Cain, M.B., O'Neill, R.P., Castillo, A.: History of optimal power flow and formulations. Federal Energy Regul. Comm. **1**, 1–36 (2012)
13. Cherukuri, A., Cortés, J.: Distributed coordination of DERs with storage for dynamic economic dispatch. IEEE Trans. Autom. Control **63**(3), 835–842 (2018)
14. Dall'Anese, E., Guggilam, S.S., Andrea Simonetto, Yu., Chen, C., Dhople, S.V.: Optimal regulation of virtual power plants. IEEE Trans. Power Syst. **33**(2), 1868–1881 (2018)
15. de Oliveira, W., Sagastizábal, C., Lemaréchal, C.: Convex proximal bundle methods in depth: a unified analysis for inexact oracles. Math. Program. **148**(1–2), 241–277 (2014)
16. Di, W., Lian, J., Sun, Y., Yang, T., Hansen, J.: Hierarchical control framework for integrated coordination between distributed energy resources and demand response. Electr. Power Syst. Res. **150**, 45–54 (2017)
17. Doan, T.T., Bose, S., Nguyen, H.D., Beck, C.L.: Convergence of the iterates in mirror descent methods. IEEE Control Syst. Lett. **3**(1), 114–119 (2019)
18. Dominguez-Garcia, A.D., Hadjicostis, C.N.: Coordination and control of distributed energy resources for provision of ancillary services. In: 2010 1st IEEE International Conference on Smart Grid Communications, pp. 537–542. IEEE, Gaithersburg, MD, USA (2010)
19. Doostizadeh, M., Aminifar, F., Lesani, H., Ghasemi, H.: Multi-area market clearing in wind-integrated interconnected power systems: a fast parallel decentralized method. Energy Convers. Manag. **113**, 131–142 (2016)

20. Duchi, J.C., Agarwal, A., Wainwright, M.J.: Dual averaging for distributed optimization: convergence analysis and network scaling. IEEE Trans. Autom. Control **57**(3), 592–606 (2012)
21. Erseghe, T.: A distributed approach to the OPF problem. EURASIP J. Adv. Signal Process. **2015**(1), 45 (2015)
22. Falsone, A., Margellos, K., Garatti, S., Prandini, M.: Dual decomposition for multi-agent distributed optimization with coupling constraints. Automatica **84**, 149–158 (2017)
23. Farivar, M., Low, S.H.: Branch flow model: relaxations and convexification—Part I & II. IEEE Trans. Power Syst. **28**(3), 2554–2572 (2013)
24. Frédéric Bonnans, J., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer, New York (2000)
25. Gan, L., Li, N., Topcu, U., Low, S.H.: Exact convex relaxation of optimal power flow in radial networks. IEEE Trans. Autom. Control **60**(1), 72–87 (2015)
26. Gharesifard, B., Basar, T., Dominguez-Garcia, A.D.: Price-based coordinated aggregation of networked distributed energy resources. IEEE Trans. Autom. Control **61**(10), 2936–2946 (2016)
27. Grangereau, M., van Ackooij, W., Gaubert, S.: Multi-stage stochastic alternating current optimal power flow with storage: bounding the relaxation gap. Electri. Power Syst. Res. **206**, 107774 (2022)
28. Guannan, Q., Li, N.: Accelerated distributed nesterov gradient descent. IEEE Trans. Autom. Control **65**(6), 2566–2581 (2020)
29. Guo, J., Hug, G., Tonguz, O.: Enabling distributed optimization in large-scale power systems (2016). arXiv:1605.09785
30. Guo, Y., Bose, S., Tong, L.: On robust tie-line scheduling in multi-area power systems. IEEE Trans. Power Syst. **33**(4), 4144–4154 (2018)
31. Gustavsson, E., Patriksson, M., Strömberg, A.-B.: Primal convergence from dual subgradient methods for convex optimization. Math. Program. **150**(2), 365–390 (2015)
32. Halilbasic, L., Chatzivasileiadis, S., Pinson, P.: Coordinating flexibility under uncertainty in multi-area AC and DC grids. In: 2017 IEEE Manchester Power Tech, pp. 1–6. IEEE, Manchester, UK (2017)
33. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
34. Huang, Y., Meng, Z., Sun, J., Ren, W.: A unified distributed method for constrained networked optimization via saddle-point dynamics. IEEE Trans. Autom. Control (2023). https://doi.org/10.1109/TAC.2023.3327940
35. Ji, Y., Tong, L.: Multi-area interchange scheduling under uncertainty. IEEE Trans. Power Syst. **33**(2), 1659–1669 (2018)
36. Ji, Y., Zheng, T., Tong, L.: Stochastic interchange scheduling in the real-time electricity market. IEEE Trans. Power Syst. **32**(3), 2017–2027 (2017)
37. Jinming, X., Tian, Y., Sun, Y., Scutari, G.: Distributed algorithms for composite optimization: unified framework and convergence analysis. IEEE Trans. Signal Process. **69**, 3555–3570 (2021)
38. Kao, H., Subramania, V.: Convergence rate analysis for distributed optimization with localization. In: 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 384–390. IEEE (2019)
39. Kim, B.H., Baldick, R.: Coarse-grained distributed optimal power flow. IEEE Trans. Power Syst. **12**(2), 932–939 (1997)
40. Lai, X., Xia, Q., Xie, L.: Inter-area power exchange preserving multi-area economic dispatch. In: 2014 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5. IEEE, National Harbor, MD, USA (2014)
41. Lai, X., Xie, L., Xia, Q., Zhong, H., Kang, C.: Decentralized multi-area economic dispatch via dynamic multiplier-based lagrangian relaxation. IEEE Trans. Power Syst. **30**(6), 3225–3233 (2015)
42. Larsson, T., Patriksson, M., Strömberg, A.-B.: Ergodic, primal convergence in dual subgradient schemes for convex programming. Math. Program. **86**(2), 283–312 (1999)
43. Lavaei, J., Low, S.H.: Zero duality gap in optimal power flow problem. IEEE Trans. Power Syst. **27**(1), 92–107 (2012)
44. Li, Z., Zeng, B., Zhang, B., Zheng, W.: Decentralized multiarea robust generation unit and tie-line scheduling under wind power uncertainty. IEEE Trans. Sustain. Energy **6**(4), 1377–1388 (2015)
45. Li, Z., Wenchuan, W., Shahidehpour, M., Zhang, B.: Adaptive robust tie-line scheduling considering wind power uncertainty for interconnected power systems. IEEE Trans. Power Syst. **31**(4), 2701–2713 (2016)
46. Li, Z., Wenchuan, W., Zeng, B., Shahidehpour, M., Zhang, B.: Decentralized contingency-constrained tie-line scheduling for multi-area power grids. IEEE Trans. Power Syst. **32**(1), 354–367 (2017)

47. Liang, S., Wang, L.Y., Yin, G.: Distributed smooth convex optimization with coupled constraints. IEEE Trans. Autom. Control **65**(1), 347–353 (2020)
48. Liang, S., Wang, L.Y., Yin, G.: Distributed dual subgradient algorithms with iterate-averaging feedback for convex optimization with coupled constraints. IEEE Trans. Cybern. **51**(5), 2529–2539 (2021)
49. Low, S.H.: Convex relaxation of optimal power flow—part II: exactness. IEEE Trans. Control Netw. Syst. **1**(2), 177–189 (2014)
50. Ma, J.: Recovery of Primal Solution in Dual Subgradient Schemes. Massachusetts Institute of Technology, Cambridge, MA (2007)
51. Madavan, A.N., Bose, S.: A stochastic primal-dual method for optimization with conditional value at risk constraints. J. Optim. Theory Appl. **190**(2), 428–460 (2021)
52. Mateos-Nunez, D., Cortes, J.: Distributed saddle-point subgradient algorithms with laplacian averaging. IEEE Transactions on Automatic Control **62**(6), 2720–2735 (2017)
53. Necoara, I., Nedelcu, V.: On linear convergence of a distributed dual gradient algorithm for linearly constrained separable convex problems. Automatica **55**, 209–216 (2015)
54. Nedić, A., Ozdaglar, A.: Approximate primal solutions and rate analysis for dual subgradient methods. SIAM J. Optim. **19**(4), 1757–1780 (2009)
55. Nedić, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control **54**(1), 48–61 (2009)
56. Nesterov, Y.: Introductory Lectures on Convex Optimization: a Basic Course. In: Applied Optimization, vol. 87. Springer, Boston, MA (2004)
57. Nesterov, Y., Shikhman, V.: Dual subgradient method with averaging for optimal resource allocation. Eur. J. Oper. Res. **270**(3), 907–916 (2018)
58. Notarnicola, I., Notarstefano, G.: Constraint-coupled distributed optimization: a relaxation and duality approach. IEEE Trans. Control Netw. Syst. **7**(1), 483–492 (2020)
59. Notarstefano, G., Notarnicola, I., Camisa, A.: Distributed optimization for smart cyber-physical networks. Found. Trends Syst. Control **7**(3), 253–383 (2020)
60. Papadaskalopoulos, D., Pudjianto, D., Strbac, G.: Decentralized coordination of microgrids with flexible demand and energy storage. IEEE Trans. Sustain. Energy **5**(4), 1406–1414 (2014)
61. Polyak, B.T.: Introduction to Optimization. Translations Series in Mathematics and Engineering. Optimization Software Publications Division, New York (1987)
62. Sherson, T., Heusdens, R., Bastiaan Kleijn, W.: On the distributed method of multipliers for separable convex optimization problems. IEEE Trans. Signal Inf. Process. Netw. **5**(3), 495–510 (2019)
63. Simonetto, A., Jamali-Rad, H.: Primal recovery from consensus-based dual decomposition for distributed convex optimization. J. Optim. Theory Appl. **168**(1), 172–197 (2016)
64. Stott, B., Jardim, J., Alsac, O.: DC power flow revisited. IEEE Trans. Power Syst. **24**(3), 1290–1300 (2009)
65. van Ackooij, W., Frangioni, A.: Incremental bundle methods using upper models. SIAM J. Optim. **28**(1), 379–410 (2018)
66. Wang, J., Zhong, H., Lai, X., Xia, Q., Shu, C., Kang, C.: Distributed real-time demand response based on lagrangian multiplier optimal selection approach. Appl. Energy **190**, 949–959 (2017)
67. Wentian, L., Liu, M., Lin, S., Li, L.: Fully decentralized optimal power flow of multi-area interconnected power systems based on distributed interior point method. IEEE Trans. Power Syst. **33**(1), 901–910 (2018)
68. Zhang, B., Tse, D.: Geometry of injection regions of power networks. IEEE Trans. Power Syst. **28**(2), 788–797 (2013)
69. Zhao Yuan and Mohammad Reza Hesamzadeh: Hierarchical coordination of TSO-DSO economic dispatch considering large-scale integration of distributed energy resources. Appl. Energy **195**, 600–615 (2017)
70. Zhao, F., Litvinov, E., Zheng, T.: A marginal equivalent decomposition method and its application to multi-area optimal power flow problems. IEEE Trans. Power Syst. **29**(1), 53–61 (2014)
71. Zhou, X., Dall'Anese, E., Chen, L., Simonetto, A.: An incentive-based online optimization framework for distribution grids. IEEE Trans. Autom. Control **63**(7), 2019–2031 (2018)

Authors and Affiliations

Haitian Liu[1] · **Subhonmesh Bose[2]** · **Hoa Dinh Nguyen[3]** · **Ye Guo[1]** · **Thinh T. Doan[4]** · **Carolyn L. Beck[2]**

Haitian Liu
liuht19@mails.tsinghua.edu.cn

Hoa Dinh Nguyen
hoa.nd@i2cner.kyushu-u.ac.jp

Ye Guo
guo-ye@sz.tsinghua.edu.cn

Thinh T. Doan
thinhdoan@vt.edu

Carolyn L. Beck
beck3@illinois.edu

[1] Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, Guangdong, China

[2] University of Illinois Urbana Champaign, Urbana, IL 61801, USA

[3] International Institute for Carbon-Neutral Energy Research (WPI-I2CNER) and Institute of Mathematics for Industry (IMI), Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

[4] Virginia Tech, Blacksburg, VA 24060, USA