A Globally Convergent Distributed Jacobi Scheme for Block-Structured Nonconvex Constrained Optimization Problems

Anirudh Subramanyam, Youngdae Kim, Michel Schanen, François Pacaud, and Mihai Anitescu, *Member, IEEE*

***Abstract*—Motivated by the increasing availability of high-performance parallel computing, we design a distributed parallel algorithm for linearly-coupled block-structured nonconvex constrained optimization problems. Our algorithm performs Jacobi-type proximal updates of the augmented Lagrangian function, requiring only local solutions of separable block nonlinear programming (NLP) problems. We provide a cheap and explicitly computable Lyapunov function that allows us to establish global and local sublinear convergence of our algorithm, its iteration complexity, as well as simple, practical and theoretically convergent rules for automatically tuning its parameters. This in contrast to existing algorithms for nonconvex constrained optimization based on the alternating direction method of multipliers that rely on at least one of the following: Gauss-Seidel or sequential updates, global solutions of NLP problems, non-computable Lyapunov functions, and hand-tuning of parameters. Numerical experiments showcase its advantages for large-scale problems, including the multi-period optimization of a 9000-bus AC optimal power flow test case over 168 time periods, solved on the Summit supercomputer using an open-source Julia code.**

***Index Terms*—distributed optimization, augmented Lagrangian, nonconvex optimization**

I. Introduction

Block-structured nonlinear optimization models are ubiquitous in science and engineering applications. Some examples include nonlinear model predictive control, multi-stage stochastic programming, supervised machine learning, optimization with differential-algebraic equations, as well as network control with geographically distributed agents. These models are used in a wide variety of areas including power systems, telecommunications, sensor networks, smart manufacturing, and chemical process systems, to name but a few. The block structure in several of these models have the following form, which is the subject of the present paper.

$$\begin{aligned} \underset{x_1,\ldots,x_T}{\text{minimize}} \quad & \sum_{t=1}^{T} f_t(x_t) \\ \text{subject to} \quad & x_t \in X_t, \quad t \in \{1,2,\ldots,T\}, \\ & \sum_{t=1}^{T} A_t x_t = b, \end{aligned} \tag{1}$$

where X_t are compact (possibly nonconvex) sets, $A_t \in \mathbb{R}^{m \times n_t}$ are matrices, and $f_t : \mathbb{R}^{n_t} \mapsto \mathbb{R}$ are continuously differentiable (possibly nonconvex) functions.

The authors are with the Mathematics and Computer Science Division, Argonne National Laboratory, Lemont IL 60439. This work was supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357.

In formulation (1), the decision variables are grouped into T blocks $(x_1, x_2, \ldots, x_T)$ coupled only via the linear constraints $\sum_{t=1}^{T} A_t x_t = b$. Such linearly coupled block-structured models arise in a wide variety of applications. For example, consider a discrete-time dynamical system:

$$x_{t+1} = \phi(x_t, u_t) \ ,$$

where $x_t \in \mathbb{R}^{n_x}$ and $u_t \in \mathbb{R}^{n_u}$ are the vectors of states and inputs, respectively, and $\phi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \mapsto \mathbb{R}^{n_x}$ is some smooth state transition function. In such systems, the optimal control problem that must be solved in the context of model predictive control [1] can be brought into the form of (1) by introducing additional variables $\tilde{x}_t \in \mathbb{R}^{n_x}$ and additional constraints:

$$x_{t+1} = \tilde{x}_t, \ \ \tilde{x}_t = \phi(x_t, u_t),$$

and by defining the t^{th} block to be $X_t = \{(x_t, u_t, \tilde{x}_t) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_x} : \tilde{x}_t = \phi(x_t, u_t)\}$. The T blocks are then coupled via the linear equations $x_{t+1} = \tilde{x}_t$, $t = 1, 2, \ldots, T-1$.

The goal of this paper is to devise a parallel algorithm that can solve instances of formulation (1) when either the number of blocks T or the number of variables/constraints in individual blocks X_t is large. In such cases, memory or storage requirements may prohibit the direct solution of (1) using conventional nonlinear programming (NLP) solvers. Also, structure-agnostic solvers may not be able to exploit any available distributed parallel computing capabilities, resulting in excessively high computational times that can be detrimental in several applications including model predictive control.

A. Literature review

A popular approach to solving large-scale instances of (1) is to decompose the problem iteratively into individual block subproblems that are easier to handle numerically. This decomposition can be done either at the linear algebra level of an interior point method [2]–[4] or by utilizing (augmented) Lagrangian functions to solve block-separable dual problems that are derived using (local) convex duality [5]–[7]. The alternating direction method of multipliers (ADMM) belongs to the latter class of methods and has seen a recent surge in popularity because of its suitability for distributed computation; e.g., see [8]. However, the vast majority of provably convergent ADMM approaches for solving (1) either exploit convexity in the objective function f_t or constraints X_t or they are tailored for specific instances of (1); e.g., see [9]–[14].

A recent body of literature [15]–[24] has analyzed the global and local convergence of ADMM approaches for nonconvex instances of formulation (1). It is well-understood by now (e.g., see [17]) that for $T > 2$, convergence requires the existence of a block variable that is constrained only by the linear coupling equation, and whose objective function is globally Lipschitz differentiable. Moreover, it must be coupled in such a way that it can "control" the iterates of the coupling constraints' dual variables (see Remark 1 later in the paper).

One way to achieve this is to first introduce an additional block of slack variables $z \in \mathbb{R}^m$ that relax the linear coupling constraints, and then drive z to 0 by adding a smooth quadratic penalty term in the objective function:

$$\begin{aligned} \underset{x_1,\ldots,x_T,z}{\text{minimize}} \;\; & \sum_{t=1}^{T} f_t(x_t) + \frac{\theta}{2}\|z\|^2 \\ \text{subject to} \;\; & z \in \mathbb{R}^m, \;\; x_t \in X_t, \;\; t \in \{1,2,\ldots,T\}, \\ & \sum_{t=1}^{T} A_t x_t + z = b, \end{aligned} \tag{2}$$

where the penalty coefficient θ is a function of the target tolerance $\epsilon > 0$ for satisfying the linear coupling constraints. In [18], this idea is exploited to design algorithms that globally converge to ϵ-stationary points of (1) in $O(\epsilon^{-6})$ iterations. Moreover, the algorithms require the global solution of NLP problems [18, Remark 3.8], which can be numerically demanding when the local subproblems are nonconvex. A similar quadratic penalty idea is adopted in [21] although it requires the augmented Lagrangian function (with respect to all nonlinear and coupling constraints) to satisfy the Kurdyka-Łojasiewicz property [15]; furthermore, that analysis does not reconcile the penalty formulation (2) with the original formulation (1). Each iteration linearizes the augmented Lagrangian function around the previous iterate that is then alternately minimized with respect to the T blocks. Other approaches using local linear and convex approximations of the nonlinear constraints or of the augmented Lagrangian function, and methods establishing convergence under additional assumptions such as coercivity of the objective function, have been proposed in [16], [23], [25], [26].

Instead of using a quadratic penalty, [19], [20] replace the objective in (2) by the augmented Lagrangian function

$$\sum_{t=1}^{T} f_t(x_t) + \beta^\top z + \frac{\theta}{2}\|z\|^2, \tag{3}$$

and drive z to 0 by iteratively updating the Lagrange multipliers β, where each so-called outer loop iteration uses another inner-level iterative ADMM scheme that minimizes (3) subject to the constraints of (2). An alternative two-level decomposition for (possibly nonlinear) coupling constraints is proposed in [22] but it depends intimately on randomized block updating of the augmented Lagrangian function which in turn requires locally tight upper bounds of the objective function with respect to each variable block, while fixing the other blocks.

All of the aforementioned algorithms rely on a Gauss-Seidel updating scheme, where the variable blocks are iteratively optimized in a particular sequence. Specifically, when updating a given variable block (say x_t), the values of all other blocks that appear earlier in the sequence (e.g., $x_1, \ldots, x_{t-1}$) must be fixed to their newest values. The overall convergence relies critically on the sequential nature of this update. Unfortunately, this makes these algorithms unsuitable for distributed parallel computations. A common workaround (e.g., see [6], [27]) to enable parallel computation is to equivalently reformulate (1), (2) by introducing additional variables y_t as follows:

$$\begin{aligned} \underset{\substack{x_1,\ldots,x_T \\ y_1,\ldots,y_T}}{\text{minimize}} \;\; & \sum_{t=1}^{T} f_t(x_t) \\ \text{subject to} \;\; & x_t \in X_t, \;\; A_t x_t = y_t, \;\; t \in \{1,2,\ldots,T\}, \\ & \sum_{t=1}^{T} y_t = b, \end{aligned} \tag{4}$$

The augmented Lagrangian function with respect to the last equation is completely decomposable in the x-variables:

$$\sum_{t=1}^{T} f_t(x_t) + \beta^\top \Big(\sum_{t=1}^{T} y_t - b \Big) + \frac{\theta}{2} \Big\| \sum_{t=1}^{T} y_t - b \Big\|^2. \tag{5}$$

Therefore, the reformulation (4) can be interpreted as a two-block model where the x- and y-variables constitute respectively the first and second blocks. Since (5) admits a closed-form minimization with respect to y (for fixed values of x), one can use any of the aforementioned Gauss-Seidel algorithms to enable parallel implementation. However, this reformulation substantially increases the problem dimension in terms of both the number of variables and constraints. This can slow down convergence and we demonstrate this empirically when we compare it with our proposed method.

Recently, [16] suggested an alternative strategy to enable parallel computation. Similar to the prox-linear method [28], it proposes to linearize around the previous iterate the quadratic penalty term in the augmented Lagrangian function, which makes the latter decomposable with respect to the x blocks. Although the method is shown to asymptotically converge for (2) (no convergence rate type is provided), it can be slow to converge in practice for models with highly nonconvex constraints. Indeed, that analysis suggests an augmented Lagrangian penalty parameter that scales as $O(\theta^4)$ posing numerical challenges. Empirically, that method failed to converge for the smallest of our test instances with $T = 3$.

B. Contributions

We propose a novel distributed Jacobi scheme for solving the block-structured optimization problem (1) with smooth nonconvex objective functions f_t and constraint sets X_t that are algebraically described by smooth nonconvex functions c_t. The scheme only requires local solutions of individual block NLP problems. In contrast to solving the higher-dimensional reformulation (4), our algorithm directly performs Jacobi updates of the augmented Lagrangian function with respect to the linear coupling constraints in (2), in which optimizing a single variable block does not require the newest values of the other blocks. This makes it particularly suitable for parallel

computation, where the cost of each iteration can be reduced by a factor of T compared to existing Gauss-Seidel schemes.

We show that the algorithm converges globally to an ϵ-approximate stationary point of (1) in no more than $O(\epsilon^{-4})$ iterations and locally converges to an approximate local minimizer at a sublinear rate under mild assumptions. The proof uses an easy-to-compute Lyapunov function that does not require a (typically unknown) local minimizer. We supplement the analysis by showing (empirically) that convergence fails if the proximal weights in the algorithm are not chosen appropriately. The algorithm can be interpreted as a nonconvex constrained extension of the Jacobi algorithms proposed in [14], [27], [29], [30], and as a Jacobi extension of the Gauss-Seidel algorithm for nonconvex problems proposed in [18]. In contrast to the majority of existing approaches for nonconvex problems, we provide a practical and automatic parameter tuning scheme, an open-source Julia implementation (that can be downloaded from https://github.com/exanauts/ProxAL.jl), and an empirical demonstration of convergence on a large-scale multi-period optimal power flow test case.

C. Notation, Assumptions and Preliminaries

For any integer T, we use $[T]$ to denote the index set $\{1, 2, \ldots, T\}$. We use x (without subscript) as shorthand for the entire vector of decisions $(x_1, \ldots, x_T) \in \mathbb{R}^n$, and X as the corresponding shorthand for the feasible set $X_1 \times \ldots \times X_T$, where we define $n \coloneqq \sum_{t=1}^{T} n_t$. For an arbitrary vector $w = (w_1, \ldots, w_T) \in \mathbb{R}^n$, we define $Aw \coloneqq \sum_{t=1}^{T} A_t w_t$, and for any $t \in [T]$, we define $A_{\neq t} w_{\neq t} \coloneqq \sum_{s \in [T] \setminus \{t\}} A_s w_s$. We let $D \coloneqq \operatorname{diag}(A_1, \ldots, A_T) \in \mathbb{R}^{Tm \times n}$ denote the block-diagonal matrix with A_1, …, A_T along its diagonal. Iterates at the k^{th} iteration are denoted with superscript k. The general normal cone [31, Definition 6.3] to a set X at a point $x \in X$ is denoted as $N_X(x)$. For a vector z and matrix M, we use $\|z\|$ and $\|M\|$ to denote their Euclidean and spectral norms, respectively. When M is positive (semi-)definite, we use $\|z\|_M$ to denote the (semi-)norm $\sqrt{z^\top M z}$. For a vector $z \in \mathbb{R}^N$ and set $S \subseteq \mathbb{R}^N$, we define $\operatorname{dist}(z, S) \coloneqq \min_{w \in S} \|z - w\|$. For any $\epsilon > 0$, we let $B_\epsilon(z)$ denote the open Euclidean ball in $\mathbb{R}^N$ with center z and radius ϵ. We use I and 0 to denote the identity and zero matrices, respectively; unless indicated otherwise, their dimensions should be clear from the context. For two square symmetric matrices M_1, M_2, we use $M_1 \succ M_2$ ($M_1 \succeq M_2$) or $M_2 \prec M_1$ ($M_2 \preceq M_2$) to indicate that $M_1 - M_2$ is positive definite (positive semidefinite).

Throughout the paper, we make the following assumptions.

(A1) X_t is non-empty and compact for all $t \in [T]$.
(A2) $f_t : \mathbb{R}^{n_t} \mapsto \mathbb{R}$ is C^2 for all $t \in [T]$.
(A3) The matrix $A \coloneqq [A_1 \ldots A_T] \in \mathbb{R}^{m \times n}$ has full row rank.
(A4) Problem (1) has a feasible solution.

We say $x^* \in X$ is a stationary point of (1), if there exist Lagrange multipliers $\lambda^* \in \mathbb{R}^m$ satisfying:

$$Ax^* = b, \tag{6a}$$
$$\nabla f_t(x_t^*) + A_t^\top \lambda^* \in -N_{X_t}(x_t^*), \quad t \in [T]. \tag{6b}$$

By introducing the primal and dual residual functions, $\pi : X \mapsto \mathbb{R}$ and $\delta_t : X_t \times \mathbb{R}^m \mapsto \mathbb{R}$, respectively, conditions (6) can be stated as $\pi(x^*) = 0$ and $\delta_t(x_t^*, \lambda^*) = 0$, $t \in [T]$, where

$$\pi(x) = \|Ax - b\|, \tag{7a}$$
$$\delta_t(x_t, \lambda) = \operatorname{dist}\left(\nabla f_t(x_t) + A_t^\top \lambda, -N_{X_t}(x_t)\right), t \in [T] \tag{7b}$$

Under an appropriate constraint qualification, the above are equivalent to the Karush-Kuhn-Tucker (KKT) conditions, and they must be necessarily satisfied if x^* is a local solution of (1); see [32, Lemma 12.9] for a discussion of the linear independence constraint qualification (LICQ).

The augmented Lagrangian function of (2) with respect to its coupling constraints is parameterized with penalty parameters $\rho, \theta > 0$ and is defined as follows (with domain $X \times \mathbb{R}^m \times \mathbb{R}^m$):

$$\begin{aligned}\mathcal{L}(x, z, \lambda) = \sum_{t=1}^{T} f_t(x_t) + \frac{\theta}{2}\|z\|^2 \\ + \lambda^\top [Ax + z - b] + \frac{\rho}{2}\|Ax + z - b\|^2.\end{aligned} \tag{8}$$

For convenience, we let $\mathcal{L}(x_t; \bar{x}_{\neq t}, \bar{z}, \bar{\lambda})$ denote the above function with domain X_t that is obtained by fixing all x_s variables to $\bar{x}_s$, $s \in [T] \setminus \{t\}$, z to $\bar{z}$, λ to $\bar{\lambda}$, and ignoring any constants:

$$\begin{aligned}\mathcal{L}(x_t; \bar{x}_{\neq t}, \bar{z}, \bar{\lambda}) = f_t(x_t) + \bar{\lambda}^\top A_t x_t \\ + \frac{\rho}{2}\|A_t x_t + A_{\neq t}\bar{x}_{\neq t} + \bar{z} - b\|^2.\end{aligned} \tag{9}$$

II. Algorithm

Algorithm 1 outlines the basic distributed scheme for solving formulation (1). Here, x^0, z^0 and λ^0 denote the initial guesses of the primal and dual variables of the penalty formulation (2), whereas ρ, θ and τ_x, τ_z are scalars denoting the penalty parameters in (2) and its augmented Lagrangian function, and proximal weights corresponding to the x and z variables, respectively.

Algorithm 1 Basic distributed proximal Jacobi scheme

Input: $x^0 \in X$, $z^0 \in \mathbb{R}^m$, $\lambda^0 \in \mathbb{R}^m$, scalars $\rho, \theta, \tau_x, \tau_z > 0$.

1: **for** $k = 1, 2, \ldots$

2: Update x: compute a local minimizer x_t^k of (10) by warm-starting with x_t^{k-1} and solve in parallel for $t \in [T]$:

$$\min_{x_t \in X_t} \mathcal{L}(x_t; x_{\neq t}^{k-1}, z^{k-1}, \lambda^{k-1}) + \frac{\tau_x}{2}\left\|x_t - x_t^{k-1}\right\|^2_{A_t^\top A_t} \tag{10}$$

3: Update z:

$$z^k = \frac{\tau_z z^{k-1} - \rho\left[Ax^k - b\right] - \lambda^{k-1}}{\tau_z + \rho + \theta} \tag{11}$$

4: Update λ:

$$\lambda^k = \lambda^{k-1} + \rho\left[Ax^k + z^k - b\right] \tag{12}$$

5: **end for**

We note a few important points about Algorithm 1. The optimization problems (10) in line 2 that compute new values of the x variables can be solved completely in parallel using any local NLP solver. These problems minimize the sum of the augmented Lagrangian function (8) and the proximal terms, over the x_t variables alone. Similarly, the update step (11) is the closed-form solution of the quadratic problem that is defined by minimizing the sum of the augmented Lagrangian function (8) and the proximal terms, over the z variables alone, and by fixing the x and λ variables to x^k and λ^{k-1}, respectively. Even though this update occurs sequentially after solving (10), our analysis and results remain unchanged if we use a variant of the algorithm where the former is performed in parallel using only information about x^{k-1}. We omit this variant for ease of exposition. Indeed, since the update step (11) is relatively cheap compared to solving problem (10), we can compute the former along with the latter locally on each node of a parallel computing architecture. Finally, the dual variables in (12) are updated as per the standard augmented Lagrangian method.

A. Global convergence

We first establish that under an appropriate choice of the scalar parameters $\rho, \theta, \tau_x, \tau_z$, the sequence $\{x^k\}$ generated by Algorithm 1 converges to a point satisfying the first-order stationarity conditions (6). The key idea is to establish that the sum of the augmented Lagrangian function and proximal terms is a candidate Lyapunov function:

$$\Phi(x,z,\lambda,\hat{x},\hat{z}) = \mathcal{L}(x,z,\lambda) + \frac{\tau_z}{4}\|z-\hat{z}\|^2 + \sum_{t=1}^{T}\frac{\tau_x}{4}\|x_t-\hat{x}_t\|^2_{A_t^\top A_t}. \tag{13}$$

Specifically, we show that the sequence $\{\Phi^k\}$, which consists of Φ evaluated at the iterates generated by the algorithm, decreases monotonically and is bounded from below by $\hat{\Phi}$:

$$\Phi^k := \Phi(x^k, z^k, \lambda^k, x^{k-1}, z^{k-1}),\ k \geq 1, \tag{14}$$

$$\hat{\Phi} := \min_{x\in X}\sum_{t=1}^{T} f_t(x_t). \tag{15}$$

In the remainder of the paper, for $k \geq 1$, we define $\Delta z^k := z^k - z^{k-1}$ and similarly define Δx^k, $\Delta\lambda^k$, and $\Delta\Phi^k$. Also, let $\Delta z^0 := -\tau_z^{-1}(\lambda^0 + \theta z^0)$, $\Delta x^0 := 0$ and $\Phi^0 := \mathcal{L}(x^0, z^0, \lambda^0) + \frac{\tau_z}{4}\|\Delta z^0\|^2$. Finally, we let $\pi^k := \pi(x^k)$ and $\delta_t^k := \delta_t(x_t^k, \lambda^k)$.

Theorem 1 (Global convergence)**.** *Suppose that assumptions (A1)–(A4) hold, and the sequence $\{(x^k, z^k, \lambda^k)\}$ is generated by Algorithm 1 with parameters $\rho, \theta, \tau_x, \tau_z > 0$, such that*

$$\eta_x := \frac{\tau_x}{4} - \frac{(T-1)\rho}{2} > 0,\ \eta_z := \frac{\tau_z}{4} - \frac{2(\theta+\tau_z)^2}{\rho} > 0. \tag{16}$$

Then, for all $K \geq 1$, there exists $j \in [K]$ such that

$$\pi^j \leq \sqrt{\frac{2\big(\Phi^1 - \hat{\Phi}\big)}{\theta}\left(1 + \frac{2(\theta+\tau_z)^2}{K\eta_z\rho}\right)}$$

$$\delta_t^j \leq (\rho+\tau_x)\|A_t\|\sqrt{\frac{2(T+1)\big(\Phi^1 - \Phi^K\big)}{K\min\{\eta_x,\eta_z\}}},\ t \in [T].$$

In particular, for any $\epsilon \in (0,1)$, if we choose

$$\theta = \frac{1}{\epsilon^2},\quad \rho = \frac{64}{\epsilon^2},\quad \tau_x = \frac{256\,(T-1)}{\epsilon^2},\quad \tau_z = \frac{2}{\epsilon^2},$$

then after $K = O(\epsilon^{-4})$ iterates, the iterates $\{(x^k, z^k, \lambda^k)\}$ converge to an ϵ-stationary point of problem (1) satisfying $\min_{j\in[K]}\max\{\pi^j, \max_{t\in[T]}\delta_t^j\} = O(\epsilon)$.

Proof. See Appendix A. □

Remark 1. *The iteration complexity can be improved to $O(\epsilon^{-2})$, whenever the original problem (1) has the same form as problem (2); see also [17]–[19] for related discussions. Specifically, we can avoid introducing variables z with penalty weights $O(\epsilon^{-2})$, whenever formulation (1) contains a variable block, say x_T, that is constrained only via the linear coupling constraints; i.e., $X_T = \mathbb{R}^{n_T}$, and its coupling matrix satisfies $A_T = \mathrm{I}$ (or more generally, its image $\mathrm{Im}\,(A_T) \supseteq \mathrm{Im}\,([b, A_1, \ldots, A_{T-1}])$), and its objective satisfies $\nabla^2 f_T(x_T) \preceq M\mathrm{I}$ for all $x \in X$, for some fixed $M > 0$.*

B. Local convergence

We now establish that the sequence $\{x^k\}$ generated by Algorithm 1 converges to an approximate local minimizer of the original problem (1), under some additional assumptions. In contrast to the previous section, where we showed global convergence to stationary points of the original problem (1), the proof proceeds by showing convergence to local minimizers of the quadratic penalty formulation (2). To relate local minimizers of the former to those of the latter, we introduce the (full) Lagrangian functions of problems (1) and (2), Λ and Λ_θ, respectively. Under assumption (A5) stated below, these functions have domains $\mathbb{R}^n\times\mathbb{R}^r\times\mathbb{R}^m$ and $\mathbb{R}^n\times\mathbb{R}^r\times\mathbb{R}^m\times\mathbb{R}^m$, respectively, where $r := \sum_{t=1}^T r_t$, and are given by:

$$\Lambda(x,\mu,\lambda) = \sum_{t=1}^{T}[f_t(x_t) + \mu_t^\top c_t(x_t)] + \lambda^\top[Ax - b] \tag{17}$$

$$\Lambda_\theta(x,z,\mu,\lambda) = \Lambda(x,\mu,\lambda) + \frac{\theta}{2}\|z\|^2 + \lambda^\top z \tag{18}$$

where $\mu_t \in \mathbb{R}^{r_t}$ and μ denotes the vector $(\mu_1, \ldots, \mu_T) \in \mathbb{R}^r$. We make the following additional assumptions.

(A5) $X_t = \{x_t \in \mathbb{R}^{n_t} : c_t(x_t) = 0\}$ can be algebraically described using only equality constraints, where $c_t : \mathbb{R}^{n_t} \mapsto \mathbb{R}^{r_t}$ is C^2 for all $t \in [T]$.

(A6) For all $\theta > 0$,[1] there exist $x^* \in X$ and $z^* \in \mathbb{R}^m$, such that (x^*, z^*) is a local solution of problem (2). Also, the constraint Jacobian $J(x^*)$ of the original problem (1) at x^* has full row rank, where we define:

$$J(x) = \begin{bmatrix}\nabla c(x)^\top \\ A\end{bmatrix}, \nabla c(x) = \begin{bmatrix}\nabla c_1(x_1) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \nabla c_T(x_T)\end{bmatrix}.$$

(A7) For all $\theta > 0$, there exist $\mu^* \in \mathbb{R}^r$ and $\lambda^* \in \mathbb{R}^m$, such that for the same x^*, z^* in assumption (A6), we have

$$w_x^\top \nabla^2_{xx}\Lambda_\theta(x^*, z^*, \mu^*, \lambda^*)w_x + \theta w_z^\top w_z > 0,$$

[1] In what follows, we suppress dependence of $x^*, z^*, \mu^*, \lambda^*$ on θ for ease of exposition.

for all $(w_x, w_z) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfying $\nabla c(x^*)^\top w_x = 0$, $Aw_x + w_z = 0$ and $\|w_x\| + \|w_z\| > 0$.

(A8) For all $k \geq 1$, the constraint Jacobian of problem (10) at its optimal solution, $\nabla c_t(x_t^k)^\top$, has full row rank.

(A9) For all $k \geq 1$, the x-update step in line 2 computes a local minimizer of problem (10) that is closest to either x_t^{k-1} if A_t has full column rank or to x_t^* otherwise.

We make some remarks about these assumptions. Assumption (A5) is without loss of generality since inequalities can be reformulated as equality constraints by adding squares of additional slack variables, and it is also used in the convergence analyses of existing augmented Lagrangian methods (e.g., see [5], [19], [24]). Assumptions (A6) and (A7) state that the linear independence constraint qualification (LICQ) and second-order sufficient conditions (SOSC) are satisfied, respectively, at the local solution (x^*, z^*) of the quadratic penalty formulation (2) for all (sufficiently large) $\theta > 0$. Assumption (A8) states that LICQ is also satisfied at local solutions of problem (10) that correspond to the x-updates. As before, these are fairly standard to establish local convergence of augmented Lagrangian methods. Finally, Assumption (A9) is satisfied by any well-behaved NLP solver if A_t has full column rank since all variables have an associated proximal term in problem (10) in this case; when A_t does not have full column rank, the assumption is necessary for convergence and it is implicit in classical analyses of the standard augmented Lagrangian method [5, see eq.(6) in Proposition 2.4].

Remark 2. *Assumption (A9) can be relaxed if the proximal term in problem* (10) *is replaced with* $\frac{\tau_x}{2}\|x - x^{k-1}\|_{P_t}^2$*, where* $P_t \succeq A_t^\top A_t$ *has full column rank. It can be shown that local (and global) convergence continues to hold in this case (with minor adjustments to the statement of Theorem 2) if we relax (A9) to require only that problem* (10) *computes a local minimizer that is closest to* x_t^{k-1}*. We do not present this generalization for ease of exposition.*

The key idea is to show that the iterates (x^k, z^k) generated by Algorithm 1 correspond to local minimizers of a perturbed variant of problem (2), that is parameterized by $p \in \mathbb{R}^m$ and $d = (d_1, \ldots, d_T, d_z) \in \mathbb{R}^{n+m}$. These parameters correspond to the primal and dual residuals of the quadratic penalty formulation (2) at the iterates of Algorithm 1. Specifically, for any $k \geq 1$, we define these as follows:

$$p^k := Ax^k + z^k - b \tag{19a}$$

$$d_t^k := \rho A_t^\top A_{\neq t} \Delta x_{\neq t}^k - \rho A_t^\top \Delta z^k - \tau_x A_t^\top A_t \Delta x_t^k \tag{19b}$$

$$d_z^k := -\tau_z \Delta z^k \tag{19c}$$

In the following theorem, recall that $D = \operatorname{diag}(A_1, \ldots, A_T)$ so that $Dx = (A_1 x_1, \ldots, A_T x_T) \in \mathbb{R}^{Tm}$.

Theorem 2 (Local convergence)**.** *Suppose that assumptions (A1)–(A9) hold,* $\epsilon \in (0, 1)$ *is a fixed constant, and parameters* $\rho, \theta, \tau_x, \tau_z > 0$ *are chosen such that* $\theta = \epsilon^{-2}$ *and* (16) *holds. Then, there exist constants* $\bar{\rho}, \epsilon' > 0$ *such that if* $\rho > \bar{\rho}$ *and the sequence* $\{(x^k, z^k, \lambda^k)\}$ *is generated by Algorithm 1 with input* (x^0, z^0, λ^0) *satisfying* $\left\|(Dx^0, z^0, \lambda^0) - (Dx^*, z^*, \lambda^*)\right\| < \epsilon'$ *and* $\lambda^0 = -\theta z^0$*, then the sequence* $\{x^k\}$ *converges at a sublinear rate to the* ϵ*-approximate local minimizer* x^* *of problem* (1) *satisfying:*

$$\max\left\{\pi(x^*), \max_{t\in[T]} \delta_t(x_t^*, \lambda^*)\right\} = O(\epsilon),$$

$$w^\top \nabla_{xx}^2 \Lambda(x^*, \mu^*, \lambda^*) w > 0, \ \forall w : J(x^*)w = 0, w \neq 0.$$

Proof. See Appendix B. □

C. Adaptive parameter tuning

In practice, the values for the parameters $\rho, \theta, \tau_x, \tau_z$ that are suggested in Theorems 1 and 2 are quite conservative and can be significantly larger than what is required for convergence. Therefore, Algorithm 2 presents a practical strategy for adaptively tuning these parameter values.

Algorithm 2 Adaptive distributed proximal Jacobi scheme

Input: $\epsilon \in (0,1)$, $\rho_0, \omega, \kappa_x, \kappa_z, \zeta, \Psi > 0$, $\nu_x, \nu_\rho, \nu_\theta, \chi > 1$

1: Initialize $\theta = \epsilon^{-2}$, $\rho = \rho_0$, $\tau_x = \kappa_x \rho$, $\tau_z = \kappa_z \rho$, $\psi = 0$.
2: **for** $k = 1, 2, \ldots$
3: Update x, z, λ as per Algorithm 1
4: **if** $\Phi^k - \Phi^{k-1} > \zeta |\Phi^k|$
5: $\tau_x \leftarrow \min\{\nu_x \tau_x, (2T-1)\rho\}$
6: **end if**
7: **if** $\max\left\{\|p^k\|_\infty, \|d^k\|_\infty\right\} \leq \epsilon$ and $\|Ax^k - b\|_\infty > \epsilon$
8: $\theta \leftarrow \nu_\theta \theta$
9: **end if**
10: **if** $\|p^k\|_\infty > \chi \|d^k\|_\infty$ and $\rho < \omega\theta$
11: $\rho \leftarrow \min\{\nu_\rho \rho, \omega\theta\}$, $\tau_x \leftarrow \kappa_x \rho$, $\tau_z \leftarrow \kappa_z \rho$
12: **else if** $\|d^k\|_\infty > \chi \|p^k\|_\infty$ and $\psi < \Psi$
13: $\rho \leftarrow \rho / \nu_\rho$, $\tau_x \leftarrow \kappa_x \rho$, $\tau_z \leftarrow \kappa_z \rho$, $\psi \leftarrow \psi + 1$
14: **end if**
15: **if** $\|Ax^k - b\|_\infty \leq \epsilon$ **stop** **end if**
16: **end for**

The parameter tuning rules are motivated from the proofs of Theorems 1 and 2. In particular, the (cheaply computable) difference in Lyapunov function values, $\Phi^k - \Phi^{k-1}$, is guaranteed to be negative, whenever the penalty parameters ρ, θ and proximal weights τ_x, τ_z are large enough to ensure $\eta_x, \eta_z > 0$. Therefore, Algorithm 2 starts with relatively small values of these parameters and gradually adjusts them. In particular, τ_x is increased in line 5 by a factor $\nu_x > 1$ whenever $\Delta\Phi^k$ (appropriately scaled by $|\Phi^k|$) is larger than some $\zeta > 0$; however, the increase is limited to $(2T-1)\rho$ based on the definition (16) of η_x. The penalty parameter θ is increased by a factor $\nu_\theta > 1$ in line 8 whenever the primal and dual residuals with respect to quadratic penalty formulation (2), p^k and d^k, are smaller than the tolerance ϵ, but the true primal residual $Ax^k - b$ continues to be larger than ϵ.

The parameter ρ is increased or decreased by a factor $\nu_\rho > 1$ to ensure that the magnitudes of p^k and d^k are within a factor $\chi > 1$ of each other, similar to [8]. However, the number of times ρ is allowed to decrease is bounded by some integer $\Psi > 0$, and moreover, it is not allowed to increase after $\rho = \omega\theta$. Therefore, ρ will not be adjusted an infinite number of times. The proximal weight τ_z is fixed at $\kappa_z \rho$, where κ_z is selected to ensure that $\eta_z > 0$ whenever ρ is also sufficiently large. In

particular, the proof of Theorem 1 shows that whenever $\rho/\theta = \omega > 32$, then $\kappa_z = 1/32$ suffices. Finally, since Theorem 1 ensures that the termination condition in line 15 will be met after a finite number of iterations (e.g., when $\eta_x, \eta_z > 0$), the parameter θ will not be increased an infinite number of times in line 8. Therefore, Algorithm 2 is guaranteed to converge to an ϵ-approximate stationary point of the original problem (1) after $O(\epsilon^{-4})$ iterations.

III. Numerical Experiments

We demonstrate the computational performance of the proximal Jacobi scheme on multi-period AC Optimal Power Flow (ACOPF) problems. The ACOPF is used to determine optimal dispatch levels of generators to satisfy electrical loads in power systems [33]. Its multi-period variant entails the solution of multiple single-period ACOPF problems that are coupled over a long time horizon. Such extended horizons are essential to model operational constraints such as ramping limits of thermal generators, as well as planning decisions involving energy storage units [34] or production costing with accurate representations of system voltages [35].

We consider a T-period ACOPF problem, where the variables $x_t = (p_t^g, q_t^g, V_t, \vartheta_t)$ in time period t are the real (p_t^g) and reactive (q_t^g) power generation levels and nodal voltage magnitudes (V_t) and angles (ϑ_t), respectively. The constraint set X_t captures AC power balances with respect to the real (p_t^d) and reactive loads (q_t^d) in time period t, whereas the objective function $f_t(p_t^g)$ minimizes the total production cost in that time period. For simplicity, we ignore constraints on transmission line flow limits. The single-period ACOPF problems are coupled via inter-temporal constraints that capture the physical ramping limits of generators; that is, generator i cannot change its real power output between periods t and $t+1$ by more than $r_i\Delta_t$ units, where Δ_t denotes the length of period t. Formally, if G and B denote the sets of generators and buses, respectively, then this problem can be formulated as follows.

$$
\begin{aligned}
\underset{x_1,\ldots,x_T}{\text{minimize}} \quad & \sum_{t\in[T]} f_t(p_t^g) \\
\text{subject to} \quad & x_t = (p_t^g, q_t^g, V_t, \vartheta_t) \in [\underline{x}, \overline{x}],\ t \in [T], \\
& \sum_{j\in G_i} p_{jt}^g - p_{it}^d = c_i^{\mathrm{re}}(V_t, \vartheta_t),\ i \in B,\ t \in [T], \\
& \sum_{j\in G_i} q_{jt}^g - q_{it}^d = c_i^{\mathrm{im}}(V_t, \vartheta_t),\ i \in B,\ t \in [T], \\
& \left|p_{i,t+1}^g - p_{it}^g\right| \le r_i\Delta_t,\ i \in G,\ t \in [T-1],
\end{aligned}
$$

where $[\underline{x}, \overline{x}]$ represent variable bounds, G_i is the set of generators connected to bus i, and c_i^{re} and c_i^{im} are smooth nonconvex functions of V_t and ϑ_t. We denote by $Y = Y^{\mathrm{re}} + \sqrt{-1}Y^{\mathrm{im}}$ the $|B| \times |B|$ admittance matrix (with $Y_{ii} = y_{ii} - \sum_{j\neq i} Y_{ij}$, $y_{ii} = y_{ii}^{\mathrm{re}} + \sqrt{-1}y_{ii}^{\mathrm{im}}$, where y_{ii}^{re} and y_{ii}^{im} are the shunt conductance and susceptance at bus $i \in B$, respectively). Using the polar formulation, the functions $c_i^{\mathrm{re,im}}$ can be written, for all buses $i \in B$, as [36]

$$
\begin{aligned}
c_i^{\mathrm{re}}(V,\vartheta) &= Y_{ii}^{\mathrm{re}} V_i^2 \\
&\quad + \sum_{j\in N_i} V_i V_j \left[Y_{ij}^{\mathrm{re}} \cos(\vartheta_i - \vartheta_j) + Y_{ij}^{\mathrm{im}} \sin(\vartheta_i - \vartheta_j)\right], \\
c_i^{\mathrm{im}}(V,\vartheta) &= -Y_{ii}^{\mathrm{im}} V_i^2 \\
&\quad + \sum_{j\in N_i} V_i V_j \left[Y_{ij}^{\mathrm{re}} \sin(\vartheta_i - \vartheta_j) - Y_{ij}^{\mathrm{im}} \cos(\vartheta_i - \vartheta_j)\right],
\end{aligned}
$$

where $N_i \subseteq B$ is the set of neighboring buses of i.

The multi-period ACOPF can be formulated as an instance of problem (1) by introducing extra variables $s_{i,t+1}^g \in [0, 2r_i\Delta_t]$ and reformulating the inter-temporal ramping limits as equality constraints: $p_{i,t+1}^g - p_{it}^g + s_{i,t+1}^g = r_i\Delta_t$. Note that although the objective function is linear or convex quadratic, the nonconvexity of c^{re} and c^{im} makes the problem nonconvex.

We consider the standard IEEE test cases '118', '1354pegase', and '9241pegase' available from MATPOWER [37]. Since the original data only provides a single vector of loads, we generate a load profile over multiple periods as follows. We first obtain hourly load data from New England ISO [38] over a typical week (7 days) of operations. We then use this $T = 24 \times 7 = 168$ period profile as a multiplier for the load at each bus, downscaling if required to ensure single-period ACOPF feasibility. Thus, the load distribution (across buses) in each period remains the same as the original MATPOWER case, and only the total load (across time periods) follows the imposed profile. We consider fairly stringent ramp limits $r_i \in [0.33\%, 0.50\%] \times p_i^{\max}$ (*per minute*, which is the standard unit for this application class), where $p_i^{\max}$ is the maximum rated output of generator i (note that $\Delta_t = 60$ min). For each test case, the objective function is scaled by 10^{-3} (roughly T^{-1}) to ensure $\rho = O(1)$. All runs are initialized with $x^0 = (\underline{x} + \overline{x})/2$, $z^0 = 0$, $\lambda^0 = 0$ and unless mentioned otherwise, following parameter values are used in Algorithm 2: $\rho_0 = 10^{-3}, \kappa_x = 2$ for case 118 and $\rho_0 = 10^{-5}, \kappa_x = 2.5$ otherwise, and $\omega = 32$, $\kappa_z = 1/32$, $\zeta = 10^{-4}$, $\nu_x = 2$, $\nu_\rho = 2$, $\nu_\theta = 10$, $\chi = 10$, $\Psi = 100$.

Our algorithm is implemented in Julia and is available at https://github.com/exanauts/ProxAL.jl. It uses the Message Passing Interface (MPI) which allows the use of distributed parallel computing resources. All NLP subproblems were solved using the JuMP modeling interface [39] and Ipopt [40] (with default options) as the NLP solver. The computational times reported in Section III-C were obtained on the Summit supercomputer at Oak Ridge National Laboratory [41].

A. Role of proximal terms

The proof of Theorem 1 shows that for fixed values of θ and ρ, the Lyapunov $\{\Phi^k\}$ must be monotonically decreasing, whenever the proximal weights τ_x are sufficiently large. To verify this empirically, we fix $\theta = 10^6$, $\rho = 1$, $\tau_z = \kappa_z\rho$ and then consider $\tau_x \in \{0, 1, 2\}$; we also disable their automatic updates in Algorithm 2. Figure 1 shows the values of the corresponding Lyapunov sequences $\{(\Phi^k - \underline{\Phi})/\underline{\Phi}\}$, where $\underline{\Phi}$ is a normalizing constant equal to the smallest observed value of Φ^k (across 100 iterations). We find that when the proximal terms are absent ($\tau_x = 0$) or when they are present but not

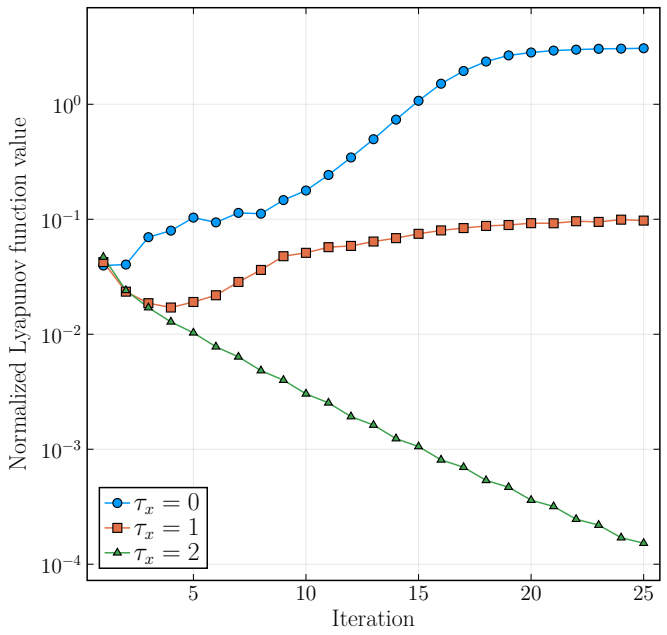

Fig. 1. Lyapunov sequence for increasing values of the proximal weight τ_x and for fixed values of θ, ρ and τ_z.

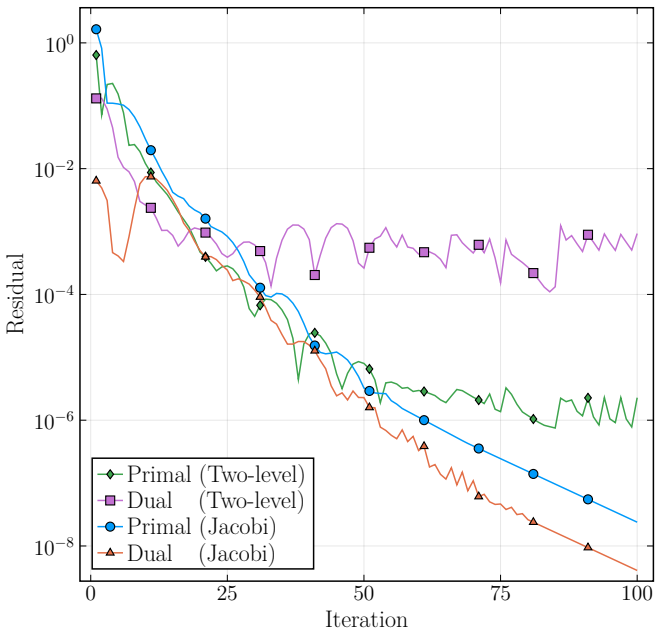

Fig. 2. Computational comparison of the proposed proximal Jacobi scheme with the two-level algorithm proposed in [19].

sufficiently large ($\tau_x = 1$), the Lyapunov sequence (and hence also $\{(x^k, z^k)\}$) diverges. This is contrast to classical Gauss-Seidel based algorithms [18], [19], which are not as critically dependent on proximal terms. Indeed, the Jacobi nature of the x-update in Algorithm 1 requires that the new iterate x^k be sufficiently close to its old value x^{k-1} for convergence.

B. Comparison with existing method

We compare the performance of our proposed scheme with the two-level ADMM algorithm proposed in [19]. Since the latter uses Gauss-Seidel or sequential updating rules, which are not amenable to distributed parallelization, the two algorithms are not directly comparable. Therefore, to solve the multi-period ACOPF with the two-level algorithm, we reformulate it using the variable splitting technique described in [19, Section 1.2]. Specifically, for each $t \in [T]$, we introduce a so-called global copy $\bar{p}_t^g$ of p_t^g, and for each $t \in [T-1]$, we introduce a so-called local copy $\hat{p}_{t+1}^g$ (local to block $t+1$) of p_t^g. The ramping limit for period $t \in [T-1]$ is then expressed as: $p_{i,t+1}^g - \hat{p}_{i,t+1}^g + s_{i,t+1}^g = r_i \Delta_t$, which is added as an additional constraint to X_{t+1}. Finally, the coupling constraints linking the various blocks can be expressed as: $p_t^g = \bar{p}_t^g$ for $t \in [T]$, and $\hat{p}_{t+1}^g = \bar{p}_t^g$ for $t \in [T-1]$. For brevity, we consider only the 118-bus case with a ramp limit of 0.33%.

As suggested in [19, Section 6], our implementation of the two-level ADMM uses $\omega = 0.75$, $\gamma = 1.5$, $[\underline{\lambda}, \bar{\lambda}] = [-10^6, 10^6]$, $\beta^1 = 1000$, $\rho^k = c_1 \beta^k$, and the k^{th} level inner-level loop is terminated based on [19, eq.14c] with $\epsilon_3^k = \sqrt{2n_g(T-1)}/(c_2 k \rho^k)$, where n_g is the number of generators and all other symbols refer to the notation in [19]. After trying $(c_1, c_2) \in \{2 \times 10^f : f = -6, -4, -2, 0, 2\} \times \{1, 1000\}$, we set $c_1 = 2 \times 10^{-4}$ and $c_2 = 1000$, since it produced the smallest infinity norm of the combined primal and dual residual vector [19, eq.14a-14c] after 50 cumulative iterations. We set the tolerance $\epsilon = 10^{-8}$ in Algorithm 2.

Figure 2 shows the performance of our proximal Jacobi scheme and the two-level algorithm. All residuals correspond to infinity norms; in our scheme, the plotted primal and dual residuals are $\|Ax^k - b\|_\infty$ and $\|d^k\|_\infty$; in the two-level scheme, they are $\max\{\|p^{g,k} - \bar{p}^{g,k}\|_\infty, \|\hat{p}^{g,k} - \bar{p}^{g,k}\|_\infty\}$ and the infinity norm of [19, eq.14a-14b] (plotted as a function of the number of cumulative iterations), respectively. We observe that for the first few iterations (up to 25) and for small target tolerances (up to 10^{-3}), the two algorithms have similar behavior in terms of reducing the optimality errors. However, our proposed scheme is able to decrease the residuals at an almost linear rate even for tighter tolerances, whereas the decrease for the two-level scheme seems to become worse than linear. This is likely because Algorithm 2 allows ρ to decrease; indeed, we observe in Figure 2 that the primal and dual residuals are roughly within a factor of $\chi = 10$ of each other at all iterations. Finally, note that the run times of both schemes are expected to be similar, since the per-iteration cost in either scheme is dominated by the solution of (single-period) ACOPF NLP problems. The actual run time is a function of the number of parallel processes available which we study in the next section.

C. Scalability

One of the main features of the proposed algorithm is its ability to make use of distributed parallel computing resources. Therefore, we study its scalability with respect to both the problem size (for a fixed number of parallel processes) as well as the number of processes (for a fixed problem size). Table I summarizes the computational performance for the different test cases using $T = 168$ parallel processes and a target tolerance of $\epsilon = 10^{-3}$ in each case. We observe from

Table I that the number of iterations and run times increase with increasing network size and with decreasing values of the ramping limit (which makes the linear coupling constraints harder to satisfy). Nevertheless, the total run time remains less than 100 minutes for the most difficult test case consisting of more than 3 million variables and constraints. Also, Figure 3 (which is plotted for the ramping limit of 0.33%) shows that although the initial primal residuals are quite large, they are reduced by more than 4 orders of magnitude over the course of roughly 50 iterations. Finally, Figure 4 shows the empirically observed run times as a function of the number of parallel processes (MPI ranks) for the 118-bus test case with $r = 0.33\%$. We find that our implementation exhibits a near-linear scaling in this particular instance.

TABLE I
SUMMARY OF COMPUTATIONAL PERFORMANCE USING $T = 168$ PARALLEL PROCESSES FOR CONVERGENCE TO $\epsilon = 10^{-3}$ TOLERANCE.

Case	# Vars.	# Cons.	Ramp %	# Iters.	Time (s)
118	66,810	48,666	0.33	24	5.2
			0.50	13	3.9
1354pegase	1,181,779	696,259	0.33	60	137.3
			0.50	53	119.7
9241pegase	3,235,756	3,148,396	0.33	67	4,755.5
			0.50	59	3,511.9

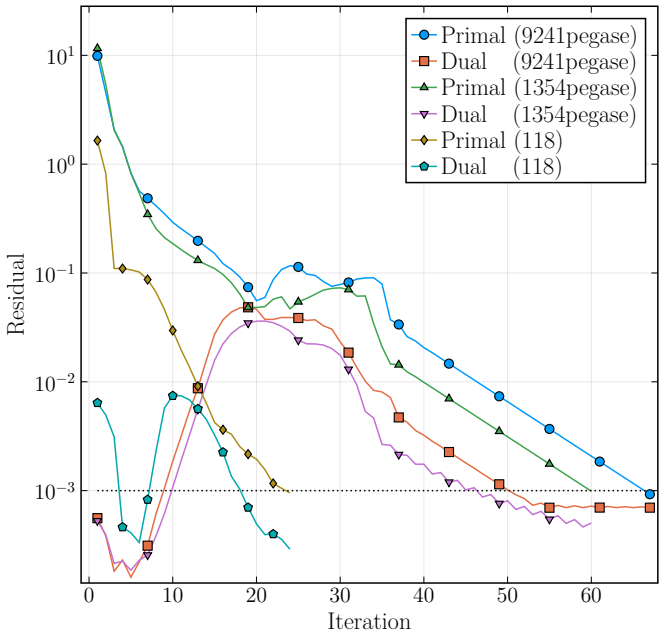

Fig. 3. Convergence of the proposed scheme to a tolerance $\epsilon = 10^{-3}$.

IV. CONCLUSIONS

This paper proposed a distributed parallel decomposition algorithm for solving linearly coupled block-structured nonconvex constrained optimization problems. Such structures naturally arise in several applications, including nonlinear model predictive control and stochastic optimization. The algorithm performs Jacobi updates of the augmented Lagrangian function with appropriately chosen proximal terms, requiring only local solutions of the individual block NLP problems. By constructing an easily computable Lyapunov function, we showed that the algorithm converges globally to ϵ-approximate stationary points in $O(\epsilon^{-4})$ iterations as well as locally to ϵ-approximate local minimizers of the original NLP at a sublinear rate. We also provided a simple, practical and theoretically convergent variant of the algorithm where its parameters are adaptively tuned during the iterations. Our numerical experiments show that it can outperform existing algorithms based on ADMM and that optimizing large-scale nonconvex AC optimal power flow problems over more than 100 time steps can be done in the order of roughly one hour. Future work includes the study of acceleration techniques such as inexact NLP solutions and momentum methods, integration with the two-level algorithm of [19], as well as the incorporation of second-order information such as in [6].

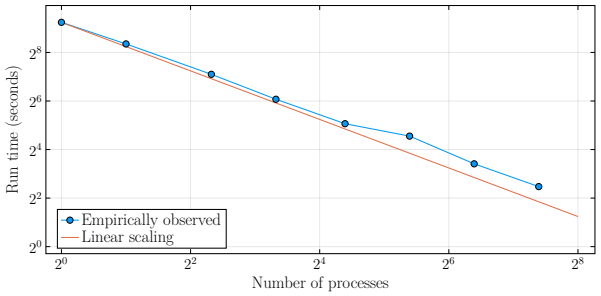

Fig. 4. Scalability as a function of the number of parallel processes.

APPENDIX A
GLOBAL CONVERGENCE PROOF

The proof of Theorem 1 requires several intermediate results. Throughout, we suppose that the conditions in Theorem 1 hold; that is, assumptions (A1)–(A4) are satisfied and $\rho, \theta, \eta_x, \eta_z > 0$ are chosen such that (16) is satisfied.

First, Lemma 2 establishes that the Lyapunov sequence $\{\Phi^k\}$ defined in (14) decreases by positive multiples of $\|\Delta x_t^k\|^2_{A_t^\top A_t}$ and $\|\Delta z^k\|^2$. Second, Lemma 3 shows that the sequence is bounded below by $\hat{\Phi}$ defined in (15). Lemma 4 uses these results to bound $\|\Delta x_t^k\|^2_{A_t^\top A_t}$ and $\|\Delta z^k\|^2$ as a function of the iteration index k. Finally, Propositions 1 and 2 establish bounds on the primal and dual residuals, respectively. The proof of Theorem 1 follows directly from these. We use the following intermediate result to simplify exposition.

Lemma 1. *For all $k \geq 0$, we have:*

$$\lambda^k = -\theta z^k - \tau_z \Delta z^k. \tag{20}$$

Proof. For $k = 0$, equation (20) is equivalent to the definition of Δz^0. For $k \geq 1$, the λ-update formula (12) implies that $\lambda^{k-1} = \lambda^k - \rho\big[\sum_{t=1}^T A_t x_t^k + z^k - b\big]$. Substituting this in the z-update formula (11) yields equation (20). □

Lemma 2. *For all $k \geq 1$, we have:*

$$\begin{aligned}\Delta\Phi^k \leq &-\eta_x \sum_{t=1}^{T} \left(\left\|\Delta x_t^k\right\|^2_{A_t^\top A_t} + \left\|\Delta x_t^{k-1}\right\|^2_{A_t^\top A_t} \right) \\ &- \eta_z \left(\left\|\Delta z^k\right\|^2 + \left\|\Delta z^{k-1}\right\|^2 \right).\end{aligned} \tag{21}$$

Proof. Using the definitions (13) and (14), we have:

$$\begin{aligned}\Delta\Phi^k = &\mathcal{L}(x^k, z^k, \lambda^k) - \mathcal{L}(x^{k-1}, z^{k-1}, \lambda^{k-1}) \\ &+ \sum_{t=1}^{T} \frac{\tau_x}{4} \left[\left\|\Delta x_t^k\right\|^2_{A_t^\top A_t} - \left\|\Delta x_t^{k-1}\right\|^2_{A_t^\top A_t} \right] \\ &+ \frac{\tau_z}{4} \left[\left\|\Delta z^k\right\|^2 - \left\|\Delta z^{k-1}\right\|^2 \right].\end{aligned} \tag{22}$$

The first term in the above expression can be written as:

$$\mathcal{L}(x^k, z^k, \lambda^k) - \mathcal{L}(x^{k-1}, z^{k-1}, \lambda^{k-1}) = (a) + (b) \tag{23a}$$
$$(a) = \left[\mathcal{L}(x^k, z^k, \lambda^k) - \mathcal{L}(x^k, z^k, \lambda^{k-1})\right] \tag{23b}$$
$$(b) = \left[\mathcal{L}(x^k, z^k, \lambda^{k-1}) - \mathcal{L}(x^{k-1}, z^{k-1}, \lambda^{k-1})\right] \tag{23c}$$

First, we examine (a). From the definitions of (8) and (12), it follows that $(a) = \frac{1}{\rho}\left\|\Delta\lambda^k\right\|^2$. Now, replacing k with $k-1$ in (20), we get $\lambda^{k-1} = -\theta z^{k-1} - \tau_z \Delta z^{k-1}$ and subtracting this equation from (20), we obtain:

$$\Delta\lambda^k = -(\theta + \tau_z)\Delta z^k + \tau_z \Delta z^{k-1}. \tag{24}$$

Combining these and using the triangle inequality, we obtain:

$$\begin{aligned}(a) &= \frac{1}{\rho}\left\|\Delta\lambda^k\right\|^2 \\ &\leq \frac{1}{\rho}\left((\theta+\tau_z)\left\|\Delta z^k\right\| + \tau_z\left\|\Delta z^{k-1}\right\|\right)^2 \\ &\leq \frac{(\theta+\tau_z)^2}{\rho}\left(\left\|\Delta z^k\right\| + \left\|\Delta z^{k-1}\right\|\right)^2 \\ &\leq \frac{2(\theta+\tau_z)^2}{\rho}\left(\left\|\Delta z^k\right\|^2 + \left\|\Delta z^{k-1}\right\|^2\right),\end{aligned} \tag{25}$$

where the second and third inequalities follow from $\theta > 0$ and $(a_1+a_2)^2 \leq 2(a_1^2+a_2^2)$ for arbitrary real a_1, a_2, respectively.

We now examine (b). For arbitrary $t \in [T]$ and $j, k, l \geq 0$, let $(x^j_{<t}, x^k_t, x^l_{>t})$ denote the vector obtained by stacking all x^j_s for $s < t$, x^k_s for $s = t$, and x^l_s for $s > t$. The term (b) can now be expressed as the following telescoping sum. Note that since λ is fixed at λ^{k-1}, we temporarily define $L(x, z) \coloneqq \mathcal{L}(x, z, \lambda^{k-1})$ to simplify exposition.

$$\begin{aligned}(b) &= \left[L(x^k, z^k) - L(x^k, z^{k-1})\right] \\ &+ \sum_{t=1}^{T}\left[L(x^k_{<t}, x^k_t, x^{k-1}_{>t}, z^{k-1}) - L(x^k_{<t}, x^{k-1}_t, x^{k-1}_{>t}, z^{k-1})\right] \\ &= \alpha + \sum_{t=1}^{T}(\beta_t + \gamma_t)\end{aligned} \tag{26}$$

where the expressions for β_t and γ_t can be verified by expanding out both sides of each summand:

$$\begin{aligned}\alpha &= L(x^k, z^k) - L(x^k, z^{k-1}) \\ &\leq -\frac{\tau_z}{2}\left\|\Delta z^k\right\|^2 \\ \beta_t &= L(x^{k-1}_{<t}, x^k_t, x^{k-1}_{>t}, z^{k-1}) - L(x^{k-1}_{<t}, x^{k-1}_t, x^{k-1}_{>t}, z^{k-1}) \\ &\leq -\frac{\tau_x}{2}\left\|\Delta x_t^k\right\|^2_{A_t^\top A_t} \\ \gamma_t &= \left[L(x^k_{<t}, x^k_t, x^{k-1}_{>t}, z^{k-1}) - L(x^{k-1}_{<t}, x^k_t, x^{k-1}_{>t}, z^{k-1})\right] \\ &\quad - \left[L(x^k_{<t}, x^{k-1}_t, x^{k-1}_{>t}, z^{k-1}) - L(x^{k-1}_{<t}, x^{k-1}_t, x^{k-1}_{>t}, z^{k-1})\right] \\ &= \rho\sum_{s=1}^{t-1}(A_t\Delta x_t^k)^\top(A_s\Delta x_s^k) \\ &\leq \frac{\rho}{2}\sum_{s=1}^{t-1}\left(\left\|A_t\Delta x_t^k\right\|^2 + \left\|A_s\Delta x_s^k\right\|^2\right) \\ \sum_{t=1}^{T}\gamma_t &\leq \frac{\rho(T-1)}{2}\sum_{t=1}^{T}\left\|\Delta x_t^k\right\|^2_{A_t^\top A_t} \\ &\leq \frac{\rho(T-1)}{2}\sum_{t=1}^{T}\left(\left\|\Delta x_t^k\right\|^2_{A_t^\top A_t} + \left\|\Delta x_t^{k-1}\right\|^2_{A_t^\top A_t}\right).\end{aligned}$$

The α-inequality is derived as follows: substituting (20) in the λ-update formula (12) yields

$$\lambda^{k-1} + \rho[Ax^k + z^k - b] + \theta z^k + \tau_z \Delta z^k = 0,$$

which is equivalent to $\nabla F(z^k) = 0$, where $F : z \mapsto \mathcal{L}(x^k, z, \lambda^{k-1}) + \frac{\tau_z}{2}\left\|z - z^{k-1}\right\|^2$ is a convex quadratic function; that is, z^k minimizes F and hence, $F(z^k) \leq F(z^{k-1})$ and this is seen to be equivalent to the α-inequality. The β_t-inequality is obtained by exploiting the fact x_t^k is locally optimal with a better objective value than x_t^{k-1} in problem (10). Finally, the γ_t-inequality follows from $a_1 a_2 \leq \frac{1}{2}(a_1^2 + a_2^2)$ for arbitrary real a_1, a_2. The inequality (21) is then obtained by combining (22), (23), (25) and (26). □

Lemma 3. *For all $k \geq 1$, we have ($\hat{\Phi}$ defined in* (15)*):*

$$\Phi^{k-1} \geq \Phi^k \geq \hat{\Phi}.$$

Proof. Since $\eta_x, \eta_z > 0$, Lemma 2 implies that $\Phi^k - \Phi^{k-1} \leq 0$ for all $k \geq 1$. Now suppose, for the sake of contradiction, that there exists $l \geq 1$ such that $\Phi^l < \hat{\Phi}$. Therefore, we must have

$$\Phi^k \leq \Phi^l < \hat{\Phi},\ \forall k \geq l. \tag{27}$$

Consider the augmented Lagrangian (8) and Lyapunov functions (13). Here, all terms except $\sum_{t=1}^{T} f_t(x_t)$ and $\lambda^\top(Ax + z - b)$ are non-negative. Therefore, using (15) and the λ-update formula (12), we have for all $j \geq 1$:

$$\begin{aligned}\Phi^j &\geq \hat{\Phi} + (\lambda^j)^\top\left(Ax^j + z^j - b\right) \\ &\geq \hat{\Phi} + \frac{1}{\rho}(\lambda^j)^\top(\lambda^j - \lambda^{j-1}) \\ &\geq \hat{\Phi} + \frac{1}{2\rho}\left(\left\|\lambda^j\right\|^2 - \left\|\lambda^{j-1}\right\|^2\right),\end{aligned}$$

where the inequality is obtained by noting that

$$2a_1^\top a_2 = \|a_1\|^2 + \|a_2\|^2 - \|a_1 - a_2\|^2 \geq \|a_1\|^2 - \|a_1 - a_2\|^2,$$

for arbitrary vectors a_1, a_2 and by setting $a_1 = \lambda^j$ and $a_2 = \lambda^j - \lambda^{j-1}$. Re-arranging the above inequality and summing over $j \in [k]$, we obtain:

$$\sum_{j=1}^{k} \big(\Phi^j - \hat{\Phi}\big) \geq -\frac{1}{2\rho} \left\|\lambda^0\right\|^2, \ \forall k \geq 1.$$

However, (27) implies:

$$\sum_{j=1}^{k} \big(\Phi^j - \hat{\Phi}\big) \leq \sum_{j=1}^{l} \big(\Phi^j - \hat{\Phi}\big) + (k-l)\big(\Phi^l - \hat{\Phi}\big), \ \forall k \geq l.$$

Since $\Phi^l - \hat{\Phi} < 0$ (by hypothesis), the right-hand side of the above inequality can be made arbitrarily smaller than 0 for sufficiently large k, contradicting the previous inequality that $\sum_{j=1}^{k} \big(\Phi^j - \hat{\Phi}\big) \geq -(2\rho)^{-1} \left\|\lambda^0\right\|^2$ for all $k \geq 1$. □

Lemma 4. *For all $K \geq 1$, there exists $j \in [K]$ such that*

$$\left\|\Delta z^j\right\|^2 + \left\|\Delta z^{j-1}\right\|^2 \leq \big(\Phi^1 - \Phi^K\big)\,(K\eta_z)^{-1}, \tag{28}$$

$$\sum_{t=1}^{T} \left\|\Delta x_t^j\right\|^2_{A_t^\top A_t} \leq \big(\Phi^1 - \Phi^K\big)\,(K\eta_x)^{-1}. \tag{29}$$

Proof. Let j be the index that maximizes the right-hand side of (21) in Lemma 2 over $k \in [K]$. The first inequality follows by summing (21) over $k \in [K]$ and noting that $\eta_x, \eta_z > 0$. The second one follows similarly if we ignore the contribution of $\|\Delta x_t^{j-1}\|_{A_t^\top A_t} \geq 0$. □

Proposition 1. *For all $K \geq 1$, there exists $j \in [K]$ such that the primal residual $\pi^j \coloneqq \pi(x^j)$, defined in* (7a)*, satisfies*

$$\pi^j \leq \sqrt{\frac{2\big(\Phi^1 - \hat{\Phi}\big)}{\theta}\left(1 + \frac{2(\theta+\tau_z)^2}{K\eta_z\rho}\right)}$$

Proof. First, we note that

$$\Phi^1 \geq \Phi^k \geq \mathcal{L}(x^k, z^k, \lambda^k) \geq \hat{\Phi} + (c), \tag{30}$$

where (c) is equal to

$$\frac{\theta}{2}\left\|z^k\right\|^2 + (\lambda^k)^\top \left[Ax^k + z^k - b\right] + \frac{\rho}{2}\left\|Ax^k + z^k - b\right\|^2.$$

Note that the first inequality in (30) follows from Lemma 3; the second follows from the definition (14) of Φ^k; and the third follows from the definitions (8) and (15) of the augmented Lagrangian function and $\hat{\Phi}$, respectively. Substituting λ^k from (20) in the expression for (c) above, we obtain

$$\begin{aligned}
(c) &= \frac{\theta}{2}\left\|z^k\right\|^2 - (\theta z^k)^\top \left[Ax^k + z^k - b\right] + \frac{\rho}{2}\left\|Ax^k + z^k - b\right\|^2 \\
&\quad - (\tau_z \Delta z^k)^\top \left[Ax^k + z^k - b\right] \\
&= \frac{\theta}{2}\left\|Ax^k - b\right\|^2 + \frac{\rho-\theta}{2}\left\|Ax^k + z^k - b\right\|^2 \\
&\quad - (\tau_z \Delta z^k)^\top \left[Ax^k + z^k - b\right] \\
&\geq \frac{\theta}{2}\big(\pi^k\big)^2 - (\tau_z \Delta z^k)^\top \left[Ax^k + z^k - b\right] \\
&= \frac{\theta}{2}\big(\pi^k\big)^2 - (\tau_z \Delta z^k)^\top \big(\frac{1}{\rho}\Delta\lambda^k\big),
\end{aligned} \tag{31}$$

where the first inequality follows from $\rho > \theta$ which can be inferred by noting that $\eta_z > 0$ implies (after completing the square in τ_z) that $(16\tau_z + 16\theta - \rho)^2 + (16\theta)^2 - (16\theta - \rho)^2 < 0$, implying $(16\theta)^2 - (16\theta - \rho)^2 < 0$, and hence, $\rho(\rho - 32\theta) > 0$, that is, $\rho > 32\theta > \theta$. Equation (31) then follows directly from the λ-update formula (12). The second term in the right-hand side of (31) can be bounded using: *(i)* the Cauchy-Schwarz inequality, *(ii)* $a_1 a_2 \leq \frac{1}{2}(a_1^2 + a_2^2)$ for real a_1, a_2, and *(iii)* relation (25) from Lemma 2, as follows:

$$\begin{aligned}
&(\tau_z \Delta z^k)^\top \big(\frac{1}{\rho}\Delta\lambda^k\big) \\
&\leq \frac{1}{2\rho}\Big(\tau_z^2 \left\|\Delta z^k\right\|^2 + \left\|\Delta\lambda^k\right\|^2\Big) \\
&\leq \frac{(\theta+\tau_z)^2}{\rho}\Big(\left\|\Delta z^k\right\|^2 + \Big(\left\|\Delta z^k\right\|^2 + \left\|\Delta z^{k-1}\right\|^2\Big)\Big).
\end{aligned} \tag{32}$$

The claim now follows from inequalities (30), (31), (32), and from the Δz^k-bound (28) established in Lemma 4. □

Proposition 2. *For all $K \geq 1$, there exists $j \in [K]$ such that the dual residual $\delta_t^j \coloneqq \delta_t(x_t^j, \lambda^j)$, defined in* (7b)*, satisfies*

$$\delta_t^j \leq (\rho + \tau_x)\left\|A_t\right\| \sqrt{\frac{(T+1)\big(\Phi^1 - \Phi^K\big)}{K \min\{\eta_x, \eta_z\}}} \tag{33}$$

for all $t \in [T]$.

Proof. Fix $t \in [T]$. Since x_t^k is locally optimal in problem (10), the first-order optimality conditions imply, for $k \geq 1$:

$$\begin{pmatrix} \nabla f_t(x_t^k) + A_t^\top \lambda^{k-1} + \tau_x A_t^\top A_t \Delta x_t^k \\ +\rho A_t^\top \Big[A_t x_t^k + A_{\neq t} x_{\neq t}^{k-1} + z^{k-1} - b\Big] \end{pmatrix} \in -N_{X_t}(x_t^k).$$

Substituting $\lambda^{k-1} = \lambda^k - \rho[Ax^k + z^k - b]$ from (12):

$$\begin{pmatrix} \nabla f_t(x_t^k) + A_t^\top \lambda^k + \tau_x A_t^\top A_t \Delta x_t^k \\ -\rho A_t^\top A_{\neq t} \Delta x_{\neq t}^k - \rho A_t^\top \Delta z^k \end{pmatrix} \in -N_{X_t}(x_t^k).$$

Definition (7b) of $\delta_t^k = \operatorname{dist}\big(\nabla f_t(x_t^k) + A_t^\top \lambda^k, -N_{X_t}(x_t^k)\big)$ implies that the latter quantity is bounded from above by:

$$\begin{aligned}
(\delta_t^k)^2 &\leq \left\|-\rho A_t^\top A_{\neq t} \Delta x_{\neq t}^k - \rho A_t^\top \Delta z^k + \tau_x A_t^\top A_t \Delta x_t^k\right\|^2 \\
&\leq \left\|A_t\right\|^2 \left(\rho \left\|A_{\neq t} \Delta x_{\neq t}^k\right\| + \rho \left\|\Delta z^k\right\| + \tau_x \left\|A_t \Delta x_t^k\right\|\right)^2 \\
&\leq (T+1)(\rho + \tau_x)^2 \left\|A_t\right\|^2 \Big(\sum_{s=1}^{T} \left\|\Delta x_s^k\right\|^2_{A_s^\top A_s} + \left\|\Delta z^k\right\|^2\Big).
\end{aligned}$$

where the inequality on the second line follows from the Cauchy-Schwarz and triangle inequalities, and the last inequality follows from $\max\{\rho, \tau_x\} \leq (\rho + \tau_x)$ and $(a_1 + \ldots + a_N)^2 \leq N(a_1^2 + \ldots a_N^2)$ for arbitrary real $a_1, \ldots, a_N$. The claim now follows from the bounds (28), (29) in Lemma 4. □

APPENDIX B
LOCAL CONVERGENCE PROOF

Throughout this section, we suppose that the conditions outlined in Theorem 2 hold. Also, for $k \geq 1$, we define $\mu^k \coloneqq (\mu_1^k, \ldots, \mu_T^k)$, where μ_t^k is the optimal Lagrange multiplier vector of the x_t-subproblem (10) at iteration k.

The key steps of the proof are as follows. We first introduce the following perturbed variant of problem (2) that is parameterized by $p \in \mathbb{R}^m$ and $d = (d_1, \ldots, d_T, d_z) \in \mathbb{R}^{n+m}$.

$$\begin{aligned} \underset{x_1,\ldots,x_T,z}{\text{minimize}} \;& \sum_{t=1}^{T} \left[f_t(x_t) - d_t^\top x_t \right] + \frac{\theta}{2} \|z\|^2 - d_z^\top z \\ \text{subject to} \;& c_t(x_t) = 0, \quad t \in [T], \\ & Ax + z = b + p. \end{aligned} \tag{$S(p,d)$}$$

Observe that $S(0,0)$ coincides precisely with problem (2). By defining p^k and d^k to be the primal and dual residuals of problem (2) at (x^k, z^k) as per eq. (19), Lemma 5 and 6 show that $(x^k, z^k, \mu^k, \lambda^k)$ must coincide precisely with the optimal solution of $S(p^k, d^k)$, whenever the former is close to $(x^*, z^*, \mu^*, \lambda^*)$ and p^k, d^k are sufficiently close to 0. This allows us to bound the difference in the augmented Lagrangian function values of problem (2) evaluated at its optimal solution and at an arbitrary iterate (Lemma 7 and 8). In Lemma 9, we establish that it is sufficient to have (Dx^k, z^k, λ^k) to be close to (Dx^*, z^*, λ^*). Finally, Proposition 3 proves that whenever the latter is true, then the sequence $\{(x^k, z^k)\}$ must converge to (x^*, z^*). This requires some intermediate results that are proved in Lemma 10 and 11. The proof of Theorem 2 follows directly from that of Proposition 3. Indeed, its first part is already true due to Theorem 1, whereas its second part follows from Assumption (A7) if we set $w_z = 0$ therein.

Lemma 5. *There exist C^1 functions $\hat{x}$, $\hat{z}$, $\hat{\mu}$, $\hat{\lambda}$ with domain $\mathbb{R}^m \times \mathbb{R}^{n+m}$ and constants $\epsilon_1, \epsilon_2 > 0$ such that*

1) $(\hat{x}(0,0), \hat{z}(0,0), \hat{\mu}(0,0), \hat{\lambda}(0,0)) = (x^*, z^*, \mu^*, \lambda^*)$.
2) *For all $(p,d) \in B_{\epsilon_1}(0)$, $(\hat{x}(p,d), \hat{z}(p,d))$ is a local minimizer of $S(p,d)$ where the LICQ and SOSC assumptions are satisfied with optimal Lagrange multipliers $(\hat{\mu}(p,d), \hat{\lambda}(p,d))$.*
3) *The functions $\hat{x}$, $\hat{z}$, $\hat{\mu}$, $\hat{\lambda}$ map to locally unique points in the sense that, for all $(p,d) \in B_{\epsilon_1}(0)$, if $(\bar{x}, \bar{z})$ is a local minimizer (or a first-order stationary point) of $S(p,d)$ with optimal Lagrange multipliers $(\bar{\mu}, \bar{\lambda})$, and if $(\bar{x}, \bar{z}, \bar{\lambda}, \bar{\mu}) \in B_{\epsilon_2}((x^*, z^*, \mu^*, \lambda^*))$, then $(\bar{x}, \bar{z}, \bar{\lambda}, \bar{\mu}) = (\hat{x}(p,d), \hat{z}(p,d), \hat{\mu}(p,d), \hat{\lambda}(p,d))$.*

Proof. The LICQ (A6) and SOSC assumptions (A7) are satisfied at (x^*, z^*), The claims then follow from a direct application of [42, Theorem 2.1] to $S(p,d)$. □

Lemma 6. *For all $k \geq 1$ such that $(p^k, d^k) \in B_{\epsilon_1}(0)$ and $(x^k, z^k, \mu^k, \lambda^k) \in B_{\epsilon_2}((x^*, z^*, \mu^*, \lambda^*))$, where ϵ_1, ϵ_2 are defined in Lemma 5 and p^k and d^k from* (19)*, we have:*

$$(\hat{x}(p^k, d^k), \hat{z}(p^k, d^k), \hat{\mu}(p^k, d^k), \hat{\lambda}(p^k, d^k)) = (x^k, z^k, \mu^k, \lambda^k).$$

Proof. We first show that (x^k, z^k, μ^k, z^k) satisfy the KKT conditions of $S(p^k, d^k)$. To see this, first note that x_t^k is locally optimal in problem (10). Under assumption (A8), this implies that it satisfies the first-order optimality condition:

$$\begin{pmatrix} \nabla f_t(x_t^k) + \nabla c_t(x_t^k)\mu_t^k + A_t^\top \lambda^{k-1} + \tau_x A_t^\top A_t \Delta x_t^k \\ \qquad + \rho A_t^\top \left[A_t x_t^k + A_{\neq t} x_{\neq t}^{k-1} + z^{k-1} - b \right] \end{pmatrix} = 0.$$

Using the λ-update (12) and eq. (19b), this is equivalent to:

$$\nabla f_t(x_t^k) + \nabla c_t(x_t^k)\mu_t^k + A_t^\top \lambda^k - d_t^k = 0.$$

Second, note that Lemma 1 along with equation (19c) implies:

$$\theta z^k + \lambda^k = d_z^k.$$

Finally, equation (19a) can be equivalently written as follows:

$$Ax^k + z^k = b + p^k.$$

Along with $c_t(x_t^k) = 0$, the last three equations are precisely the KKT conditions of $S(p^k, d^k)$. The result now follows from the third part of Lemma 5. □

Lemma 7. *There exist constants $\rho_1, \epsilon_3 > 0$ such that for all $\rho > \rho_1$ and $k \geq 1$ where $(x^k, z^k) \in B_{\epsilon_3}((x^*, z^*))$, we have:*

$$\begin{aligned} &\sum_{t\in[T]} f_t(x_t^k) + \frac{\theta}{2} \left\|z^k\right\|^2 + (\lambda^*)^\top p^k + \frac{\rho}{4} \left\|p^k\right\|^2 \\ &\quad \geq \sum_{t\in[T]} f_t(x_t^*) + \frac{\theta}{2} \left\|z^*\right\|^2. \end{aligned} \tag{34}$$

Proof. Define the (fully) augmented Lagrangian function of $S(0,0)$ with respect to the optimal multipliers (μ^*, λ^*), as follows $\mathcal{S} : \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}$, where $\mathcal{S}(x,z)$ is given by:

$$\Lambda_\theta(x, z, \mu^*, \lambda^*) + \frac{\rho}{4} \sum_{t=1}^{T} \|c_t(x_t)\|^2 + \frac{\rho}{4} \|Ax + z - b\|^2$$

The definition of (x^*, z^*) as a local minimizer of problem $S(0,0)$ satisfying the LICQ assumption, means that $\nabla \mathcal{S}(x^*, z^*) = 0$ and $c_t(x_t^*) = 0$. The latter also implies that the Hessian $\nabla^2 \mathcal{S}(x^*, z^*)$ is given by:

$$\begin{bmatrix} \nabla^2_{xx} \Lambda(x^*, z^*, \mu^*, \lambda^*) & 0 \\ 0 & \theta \mathrm{I} \end{bmatrix} + \frac{\rho}{2} \begin{bmatrix} J(x^*)^\top J(x^*) & A^\top \\ A & \mathrm{I} \end{bmatrix}.$$

Assumption (A7) and [43, Lemma 3.2.1] ensure the existence of $\rho_1 > 0$ such that $\nabla^2 \mathcal{S}(x^*, z^*)$ is positive definite and that its minimum eigenvalue is larger than some fixed $\epsilon_3' > 0$, for all $\rho > \rho_1$. Continuity of $\nabla^2 \mathcal{S}$ along with the second-order sufficient conditions for unconstrained minimization then imply the existence of $\epsilon_3 > 0$ such that (x^*, z^*) is a local minimizer of $\mathcal{S}$ in some neighborhood of radius ϵ_3 around (x^*, z^*). Note that the radius ϵ_3 is independent of ρ as long as the minimum eigenvalue of $\nabla^2 \mathcal{S}(x^*, z^*)$ remains larger than $\epsilon_3' > 0$. The statement of the lemma then follows by noting that $\mathcal{S}(x^k, z^k) \geq \mathcal{S}(x^*, z^*)$ and after substituting the expressions for $\mathcal{S}(x,z)$ and for p^k from (19a). □

Lemma 8. *There exist constants $\rho_2, \epsilon_4 > 0$ such that for all $\rho > \rho_2$ and $k \geq 1$ where $(p^k, d^k) \in B_{\epsilon_4}(0)$ and $(x^k, z^k, \mu^k, \lambda^k) \in B_{\epsilon_2}((x^*, z^*, \mu^*, \lambda^*))$, with ϵ_2 defined in Lemma 5, we have:*

$$\begin{aligned} &\sum_{t\in[T]} \left(f_t(x_t^k) - d_t^k x_t^k\right) + \frac{\theta}{2} \left\|z^k\right\|^2 - (d_z^k)^\top z^k + (\lambda^k)^\top p^k - \frac{\rho}{4} \left\|p^k\right\|^2 \\ &\quad \leq \sum_{t\in[T]} \left(f_t(x_t^*) - d_t^k x_t^*\right) + \frac{\theta}{2} \left\|z^*\right\|^2 - (d_z^k)^\top z^*. \end{aligned} \tag{35}$$

Proof. Define the primal functional corresponding to $S(p,d)$, as follows $\mathcal{Q}:\mathbb{R}^m\times\mathbb{R}^{n+m}\mapsto\mathbb{R}$, where $\mathcal{Q}(p,d)$ is given by:

$$\sum_{t=1}^{T}\left[f_t(\hat{x}_t(p,d))-d_t^\top\hat{x}_t(p,d)\right]+\frac{\theta}{2}\left\|\hat{z}(p,d)\right\|^2-d_z^\top\hat{z}(p,d)\ .$$

Now [5, Proposition 1.28] implies that $\nabla_p Q(p,d)=-\hat{\lambda}(p,d)$, which is continuously differentiable from Lemma 5. Therefore, $\nabla^2_{pp}Q$ is continuous and [43, Lemma 3.2.1] ensures the existence of $\rho_2>0$ such that $\nabla^2_{pp}Q(0,0)+\frac{\rho}{2}\mathrm{I}$ is positive definite and that its minimum eigenvalue is bounded strictly away from 0 for all $\rho>\rho_2$. By a similar argument as in the proof of Lemma 7, continuity of $\nabla^2_{pp}Q$ implies that there exists $\epsilon_4'>0$ (independent of ρ) such that $\nabla^2_{pp}Q(p,d)+\frac{\rho}{2}\mathrm{I}$ is positive definite, and hence the function $F_d:p\mapsto Q(p,d)+\frac{\rho}{4}\left\|p\right\|^2$ is convex, whenever $(p,d)\in B_{\epsilon_4'}(p,d)$. Now fix $\epsilon_4=\min\{\epsilon_4',\epsilon_1\}>0$ and $d=d^k$. Convexity of F_{d_k} implies:

$$F_{d_k}(p^k)+(0-p^k)^\top\nabla F_{d_k}(p^k)\leq F_{d_k}(0).$$

The statement of the lemma now follows by *(i)* substituting $F_{d_k}(p^k)=Q(p^k,d^k)+\frac{\rho}{4}\left\|p^k\right\|^2$, replacing $Q(p^k,d^k)$ using its definition, and noting $(\hat{x}(p^k,d^k),\hat{z}(p^k,d^k))=(x^k,z^k)$ from Lemma 6; *(ii)* substituting $\nabla F_{d_k}(p^k)=-\hat{\lambda}(p^k,d^k)+\frac{\rho}{2}p^k=-\lambda^k+\frac{\rho}{2}p^k$ (from Lemma 6); and, *(iii)* $F_{d_k}(0)=Q(0,d^k)$ is less than or equal to the right-hand side of (35), since (x^*,z^*) is feasible (but possibly suboptimal) in $S(0,d^k)$. □

Lemma 9. *There exist $\rho_3,\epsilon_5>0$ such that for all $\rho>\rho_3$ and $k\geq 1$ where $(Dx^{k-1},z^{k-1},\lambda^{k-1})\in B_{\epsilon_5}((Dx^*,z^*,\lambda^*))$, we have $(x^k,z^k,\mu^k,\lambda^k)\in B_{\min\{\epsilon_2,\epsilon_3\}}((x^*,z^*,\mu^*,\lambda^*))$, where ϵ_2,ϵ_3 are defined in Lemma 5 and 7.*

Proof. Observe that for fixed choices of $\rho,\theta,\tau_x,\tau_z$ in Algorithm 1, the x-update step (10) at iteration k depends only on the values of Dx^{k-1}, z^{k-1} and λ^{k-1} at the previous iteration; therefore, we can define $G:(Dx^{k-1},z^{k-1},\lambda^{k-1})\mapsto(x^k,\mu^k)$ to be the corresponding mapping. Observe now that the claim follows trivially if we can show that G is continuous and satisfies $G(Dx^*,z^*,\lambda^*)=(x^*,\mu^*)$. Indeed, if this is true, then observe that the z-update (11) and λ-update (12) formulas also have the same properties: they define continuous maps and satisfy $z^k=z^*$ and $\lambda^k=\lambda^*$ whenever $x^k=x^*$, $z^{k-1}=z^*$ and $\lambda^{k-1}=\lambda^*$, since $Ax^*+z^*=b$ and $\lambda^*=-\theta z^*$ (from the KKT conditions of problem (2)).

Now fix $t\in[T]$ and $(Dx^{k-1},z^{k-1},\lambda^{k-1})=(Dx^*,z^*,\lambda^*)$. Observe that problem (10) satisfies the LICQ assumption at x_t^*, because of (A6). Moreover, since $Ax^*+z^*=b$, observe that (x_t^*,μ_t^*) is also a first-order stationary point of problem (10):

$$\begin{pmatrix}\nabla f_t(x_t^*)+\nabla c_t(x_t^*)\mu_t^*+A_t^\top\lambda^*+\tau_x A_t^\top A_t(x_t^*-x_t^*)\\ +\rho A_t^\top\left[A_tx_t^*+A_{\neq t}x_{\neq t}^*+z^*-b\right]\end{pmatrix}=0.$$

To show that it satisfies SOSC, note that Assumption (A7) and [43, Lemma 3.2.1] ensure the existence of $\rho_3>0$ such that

$$\begin{bmatrix}\nabla^2_{xx}\Lambda(x^*,z^*,\mu^*,\lambda^*) & 0\\ 0 & \theta\mathrm{I}\end{bmatrix}+\rho\begin{bmatrix}A^\top A & A^\top\\ A & \mathrm{I}\end{bmatrix}.$$

is positive definite on the domain $\{(w_x,w_z)\in\mathbb{R}^n\times\mathbb{R}^m\setminus\{0\}:\nabla c(x^*)^\top w_x=0\}$ for all $\rho>\rho_3$. This means that for $w_{x_t}\neq 0$, $\nabla c_t(x_t^*)^\top w_{x_t}=0$, $w_z=0$, and $w_{x_s}=0$ for $s\neq t$, we have:

$$w_{x_t}^\top\left[\nabla^2_{x_tx_t}\Lambda(x^*,z^*,\mu^*,\lambda^*)+(\rho+\tau_x)A_t^\top A_t\right]w_{x_t}>0,$$

which is precisely the SOSC condition for problem (10) at (x_t^*,μ_t^*), and therefore, (x^*,μ^*) is a strict local solution.

Continuity of the mapping G now follows directly from Assumption (A9) and classical NLP sensitivity [42, Theorem 2.1] applied to problem (10) with Dx_t^{k-1}, z^{k-1} and λ^{k-1} viewed as perturbation parameters. □

Lemma 10. *The sequences $\{p^k\}$ and $\{d^k\}$ converge to 0.*

Proof. After summing inequality (22) over $k\in[K]$ and noting $\Phi^K\geq\hat{\Phi}$ from Lemma 3, we obtain that the sequences $\{\Delta z^k\}$ and $\{A_t\Delta x_t^k\}$ for $t\in[T]$ must all converge to 0. Therefore, the definitions (19b) and (19c) imply that $\{d^k\}$ also converges to 0. Equations (12) and (24) imply $p^k=\frac{1}{\rho}\Delta\lambda^k=-\frac{\theta+\tau_z}{\rho}\Delta z^k+\frac{\tau_z}{\rho}\Delta z^{k-1}$, which means that the sequence $\{p^k\}$ must also converge to 0. □

Lemma 11. *1) Any two vectors a_1,a_2 of equal dimension satisfy $2a_1^\top a_2=\left\|a_1+a_2\right\|^2-\left\|a_1\right\|^2-\left\|a_2\right\|^2$.*

2) For any $w=(w_1,\ldots,w_T)\in\mathbb{R}^{n_1}\times\ldots\times\mathbb{R}^{n_T}$, we have:

$$\begin{aligned}&(\rho+\tau_x)\sum_{t=1}^{T}\left\|w_t\right\|^2_{A_t^\top A_t}-\rho\left\|w\right\|^2_{A^\top A}=\left\|Dw\right\|^2_R,\\ &R\coloneqq(\rho+\tau_x)\mathrm{I}-\rho EE^\top\succ 0,\\ &E^\top\coloneqq\begin{bmatrix}\mathrm{I}_{m\times m} & \ldots & \mathrm{I}_{m\times m}\end{bmatrix}\in\mathbb{R}^{m\times Tm}.\end{aligned}$$

Proof. The first part follows by expanding its right-hand side. To prove the second part, note that $D=\mathrm{diag}(A_1,\ldots,A_T)$ can be used to verify that $\sum_{t=1}^T\left\|w_t\right\|^2_{A_t^\top A_t}=w^\top D^\top Dw$. Similarly, the definition of E can be used to verify that $A=E^\top D$ and hence, $\left\|w\right\|^2_{A^\top A}=w^\top D^\top EE^\top Dw$. This proves the claimed equation. To show that $R\succ 0$, observe that $E\ E^\top$ can be equivalently expressed as the Kronecker product $\mathrm{e}\,\mathrm{e}^\top\otimes\mathrm{I}_{m\times m}$, where e is the vector of ones in $\mathbb{R}^T$. Therefore, its eigenvalues are given by pairwise products of the eigenvalues of $\mathrm{e}\,\mathrm{e}^\top$ (which are 0 and T) and $\mathrm{I}_{m\times m}$ (which is 1). Therefore, the eigenvalues of $E\ E^\top$ are 0 and T, and hence, those of $(\rho+\tau_x)\mathrm{I}-\rho E\ E^\top$ are $\rho+\tau_x$ and $\rho+\tau_x-\rho T$, both of which are positive since $\eta_x>0$, see (16). □

Proposition 3. *There exist constants $\bar{\rho},\epsilon'>0$ such that if $\rho>\bar{\rho}$ and $\left\|(Dx^0,z^0,\lambda^0)-(Dx^*,z^*,\lambda^*)\right\|<\epsilon'$, then the sequence $\{(x^k,z^k,\lambda^k,\Delta z^k)\}$ converges to $(x^*,z^*,\lambda^*,0)$ at a sublinear rate.*

Proof. Set $\bar{\rho}=\max\{\rho_1,\rho_2,\rho_3\}$, where the latter are defined in Lemma 7, 8, 9. Also, assume for the moment that the other conditions of Lemma 7 and 8 are also satisfied for all $k\geq 1$;

we shall shortly show how ϵ' can be chosen to ensure this. Adding the corresponding equations (34) and (35), we obtain:

$$\underbrace{(\lambda^* - \lambda^k)^\top p^k + \frac{\rho}{2}\left\|p^k\right\|^2}_{(d)} + \underbrace{(d_z^k)^\top (z^k - z^*)}_{(e)} + \underbrace{\sum_{t=1}^{T}(d_t^k)^\top (x_t^k - x_t^*)}_{(f)} \geq 0.$$

We now we have the following

$$\begin{aligned}(d) &= \frac{1}{\rho}(\lambda^* - \lambda^k)^\top \Delta\lambda^k + \frac{1}{2\rho}\left\|\Delta\lambda^k\right\|^2 \\ &= \frac{1}{2\rho}\big[\left\|\lambda^* - \lambda^{k-1}\right\|^2 - \left\|\lambda^* - \lambda^k\right\|^2\big],\end{aligned}$$

where the first equality follows from definition (19a) and the λ-update formula (12), and the second equality follows from the first part of Lemma 11. Subsequently,

$$\begin{aligned}(e) &= \tau_z(\Delta z^k)^\top (z^* - z^k) \\ &= \frac{\tau_z}{2}\big[\left\|z^* - z^{k-1}\right\|^2 - \left\|z^* - z^k\right\|^2 - \left\|\Delta z^k\right\|^2\big],\end{aligned}$$

where again the first equality follows from definition (19c) and the second follows from the first part of Lemma 11. Finally,

$$\begin{aligned}(f) &= \sum_{t=1}^{T}\left[\rho A\Delta x^k + \rho\Delta z^k - (\rho+\tau_x)A_t\Delta x_t^k\right]^\top A_t(x_t^k - x_t^*) \\ &= (f_1) + (f_2),\end{aligned}$$

where the first equality follows from definition (19b) and after adding and subtracting $\rho A_t\Delta x_t^k$ from the latter. For the first component of (f) we have

$$\begin{aligned}(f_1) &= -\rho(A\Delta x^k)^\top A(x^* - x^k) \\ &\quad + (\rho+\tau_x)\sum_{t=1}^{T}(A_t\Delta x_t^k)^\top A_t(x_t^* - x_t^k) \\ &= -\frac{\rho}{2}\left[\left\|x_t^* - x_t^{k-1}\right\|_{A^\top A}^2 - \left\|x_t^* - x_t^k\right\|_{A^\top A}^2 - \left\|\Delta x_t^k\right\|_{A^\top A}^2\right] \\ &\quad + \frac{\rho+\tau_x}{2}\sum_{t=1}^{T}\left[\begin{array}{l}\left\|x_t^* - x_t^{k-1}\right\|_{A_t^\top A_t}^2 - \left\|x_t^* - x_t^k\right\|_{A_t^\top A_t}^2 \\ - \left\|\Delta x_t^k\right\|_{A_t^\top A_t}^2\end{array}\right] \\ &= \frac{1}{2}\big[\left\|D(x^* - x^{k-1})\right\|_R^2 - \left\|D(x^* - x^k)\right\|_R^2 - \left\|D\Delta x^k\right\|_R^2\big]\end{aligned}$$

where the second and third equalities follow from the first and second parts of Lemma 11, respectively. For the second component of f we have

$$\begin{aligned}(f_2) &= \rho(\Delta z^k)^\top A(x^k - x^*) \\ &= (\Delta z^k)^\top\Delta\lambda^k + \rho(\Delta z^k)^\top (z^* - z^k) \\ &= -(\theta+\tau_z)\left\|\Delta z^k\right\|^2 + \tau_z(\Delta z^k)^\top\Delta z^{k-1} \\ &\quad + \frac{\rho}{2}\big[\left\|z^* - z^{k-1}\right\|^2 - \left\|z^* - z^k\right\|^2 - \left\|\Delta z^k\right\|^2\big] \\ &\leq \frac{\rho}{2}\big[\left\|z^* - z^{k-1}\right\|^2 - \left\|z^* - z^k\right\|^2\big] + \frac{\tau_z}{2}\left\|\Delta z^k\right\|^2,\end{aligned}$$

where the second equality follows by substituting $Ax^* = b - z^*$ and $Ax^k = b - z^k + \frac{1}{\rho}\Delta\lambda^k$, the third equality follows from (24) and the first part of Lemma 11, and the inequality follows by noting first that $2(\Delta z^k)^\top\Delta z^{k-1} \leq \|\Delta z^k\|^2 + \|\Delta z^{k-1}\|^2$ and then that $\theta + (\tau_z/2) + \rho > 0$.

Combining the equations for $(d), (e), (f), (f_1)$, the inequality for (f_2), and $(d)+(e)+(f) \geq 0$, we obtain:

$$\begin{aligned}&\left\|(Dx^k, z^k, \lambda^k) - (Dx^*, z^*, \lambda^*)\right\|_*^2 + \tau_z\left\|\Delta z^k\right\|^2 + \left\|D\Delta x^k\right\|_R^2 \\ &\leq \left\|(Dx^{k-1}, z^{k-1}, \lambda^{k-1}) - (Dx^*, z^*, \lambda^*)\right\|_*^2 + \tau_z\left\|\Delta z^{k-1}\right\|^2,\end{aligned} \tag{36}$$

where we define the norm:

$$\|(Dx, z, \lambda)\|_* := \sqrt{\|Dx\|_R^2 + (\rho+\tau_z)\|z\|^2 + (1/\rho)\|\lambda\|^2}.$$

Summing inequality (36) over $k \in [j]$ and using $\Delta z^0 = (-\lambda^0 - \theta z^0)/\tau_z = 0$ (by hypothesis), we obtain for all $j \geq 1$:

$$\begin{aligned}&\left\|(Dx^j, z^j, \lambda^j) - (Dx^*, z^*, \lambda^*)\right\|_*^2 + \tau_z\left\|\Delta z^j\right\|^2 + \left\|D\Delta x^j\right\|_R^2 \\ &\leq \left\|(Dx^0, z^0, \lambda^0) - (Dx^*, z^*, \lambda^*)\right\|_*^2\end{aligned}$$

The equivalence of norms implies that whenever the Euclidean norm $\left\|(Dx^0, z^0, \lambda^0) - (Dx^*, z^*, \lambda^*)\right\|$ is less than ϵ', the right-hand side of the above inequality is also sufficiently small, and the terms on the left-hand side are even smaller. In particular, by choosing a small ϵ', we can ensure: *(i)* (Dx^j, z^j, λ^j) remains close to (Dx^*, z^*, λ^*); and *(ii)* $D\Delta x^j$ and Δz^j are close to 0, which implies p^j, d^j are also close to 0 (see argument in proof of Lemma 10). Finally, note that *(i)* and *(ii)* satisfy the conditions of Lemma 9 (which in turn allows us to satisfy those of Lemma 7) and Lemma 8, respectively.

To be precise, the equivalence of norms implies there exist $0 < c_1 \leq c_2$ such that $c_1\|\cdot\|_* \leq \|\cdot\| \leq c_2\|\cdot\|_*$. Now set $\epsilon' = \min\{\frac{c_1}{c_0}\epsilon_1, \frac{c_1}{c_2}\epsilon_5\} < \epsilon_5$, where ϵ_1 and ϵ_5 are defined in Lemma 5 and 9, respectively, and $c_0 := 4(\rho + \theta + \tau_x + \tau_z + 1)(\frac{1}{\rho}+1)(\frac{1}{\sqrt{\tau_z}}+1)(T+1)(\|D\|+1)(c_2+1)$ is sufficiently large. Then, it can be verified that for all $k \geq 0$, we have: $(Dx^k, z^k, \lambda^k) \in B_{\epsilon_5}(Dx^*, z^*, \lambda^*)$, which verifies the conditions of Lemma 9 and hence of Lemma 7, and $(p^k, d^k) \in B_{\epsilon_1}(0)$, which verifies the condition of Lemma 8. Now, since the conditions of Lemma 6 are also satisfied, and since $\{p^k\}$ and $\{d^k\}$ converge to 0 (from Lemma 10), this proves that $\{(x^k, z^k, \lambda^k)\} = \{(\hat{x}(p^k, d^k), \hat{z}(p^k, d^k), \hat{\lambda}(p^k, d^k))\}$ converge to $(\hat{x}(0,0), \hat{z}(0,0), \hat{\lambda}(0,0)) = (x^*, z^*, \lambda^*)$ from Lemma 5. Finally, the convergence rate follows from (36) as follows:

$$\frac{\left\|(Dx^k, z^k, \lambda^k, \Delta z^k) - (Dx^*, z^*, \lambda^*, 0)\right\|_\dagger^2}{\left\|(Dx^{k-1}, z^{k-1}, \lambda^{k-1}, \Delta z^{k-1}) - (Dx^*, z^*, \lambda^*, 0)\right\|_\dagger^2} \leq 1,$$

where we have defined the norm:

$$\|(Dx, z, \lambda, \Delta z)\|_\dagger := \sqrt{\|(Dx, z, \lambda)\|_*^2 + \tau_z\|\Delta z\|^2}.$$

The above is equivalent to the definition of Q-sublinear convergence [5, Proposition 1.1d] which proves the claim. □

REFERENCES

[1] L. Grüne and J. Pannek, *Nonlinear Model Predictive Control*. Springer, 2017.

[2] J. Gondzio and A. Grothey, "Exploiting structure in parallel implementation of interior point methods for optimization," *Computational Management Science*, vol. 6, no. 2, pp. 135–160, 2009.

[3] N. Chiang, C. G. Petra, and V. M. Zavala, "Structured nonconvex optimization of large-scale energy systems using PIPS-NLP," in *2014 Power Systems Computation Conference*. IEEE, 2014, pp. 1–7.

[4] D. Kourounis, A. Fuchs, and O. Schenk, "Toward the next generation of multiperiod optimal power flow solvers," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4005–4014, 2018.

[5] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[6] B. Houska, J. Frasch, and M. Diehl, "An augmented lagrangian based algorithm for distributed nonconvex optimization," *SIAM Journal on Optimization*, vol. 26, no. 2, pp. 1101–1127, 2016.

[7] J.-H. Hours and C. N. Jones, "A parametric nonconvex decomposition algorithm for real-time and distributed NMPC," *IEEE Transactions on Automatic Control*, vol. 61, no. 2, pp. 287–302, 2016.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[9] G. Li and T. K. Pong, "Global convergence of splitting methods for nonconvex composite optimization," *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2434–2460, 2015.

[10] T. Lin, S. Ma, and S. Zhang, "On the global linear convergence of the ADMM with multiblock variables," *SIAM Journal on Optimization*, vol. 25, no. 3, pp. 1478–1497, 2015.

[11] S. Magnússon, P. C. Weeraddana, M. G. Rabbat, and C. Fischione, "On the Convergence of Alternating Direction Lagrangian Methods for Nonconvex Structured Optimization Problems," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 3, pp. 296–309, 2016.

[12] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.

[13] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers," *Mathematical Programming*, vol. 162, no. 1-2, pp. 165–199, 2017.

[14] J. G. Melo and R. D. Monteiro, "Iteration-complexity of a Jacobi-type non-Euclidean ADMM for multi-block linearly constrained nonconvex programs," *arXiv preprint arXiv:1705.07229*, 2017.

[15] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1, pp. 459–494, 2014.

[16] Q. Liu, X. Shen, and Y. Gu, "Linearized ADMM for nonconvex nonsmooth optimization with convergence analysis," *IEEE Access*, vol. 7, pp. 76 131–76 144, 2019.

[17] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, 2019.

[18] B. Jiang, T. Lin, S. Ma, and S. Zhang, "Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis," *Computational Optimization and Applications*, vol. 72, no. 1, pp. 115–157, 2019.

[19] K. Sun and X. A. Sun, "A two-level distributed algorithm for nonconvex constrained optimization," *arXiv preprint arXiv:1902.07654*, 2019.

[20] W. Tang and P. Daoutidis, "Fast and stable nonconvex constrained distributed optimization: the ELLADA algorithm," *Optimization and Engineering*, pp. 1–43, 2021.

[21] Y. Yang, G. Hu, and C. J. Spanos, "A proximal linearization-based decentralized method for nonconvex problems with nonlinear constraints," *arXiv preprint arXiv:2001.00767*, 2020.

[22] Q. Shi and M. Hong, "Penalty Dual Decomposition Method for Non-smooth Nonconvex Optimization—Part I: Algorithms and Convergence Analysis," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4108–4122, 2020.

[23] D. Zhu, L. Zhao, and S. Zhang, "A first-order primal-dual method for nonconvex constrained optimization based on the augmented lagrangian," *arXiv preprint arXiv:2007.12219*, 2020.

[24] S. M. Harwood, "Analysis of the alternating direction method of multipliers for nonconvex problems," in *SN Operations Research Forum*, vol. 2, no. 1. Springer, 2021, pp. 1–29.

[25] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization—part i: Theory," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 1929–1944, 2016.

[26] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Linear convergence of first-and zeroth-order primal-dual algorithms for distributed nonconvex optimization," *arXiv preprint arXiv:1912.12110*, 2019.

[27] W. Deng, M.-J. Lai, Z. Peng, and W. Yin, "Parallel multi-block ADMM with $o(1/k)$ convergence," *Journal of Scientific Computing*, vol. 71, no. 2, pp. 712–736, 2017.

[28] G. Chen and M. Teboulle, "A proximal-based decomposition method for convex minimization problems," *Mathematical Programming*, vol. 64, no. 1, pp. 81–101, 1994.

[29] G. Banjac, K. Margellos, and P. J. Goulart, "On the Convergence of a Regularized Jacobi Algorithm for Convex Optimization," *IEEE Transactions on Automatic Control*, vol. 63, no. 4, pp. 1113–1119, 2018.

[30] N. Chatzipanagiotis and M. M. Zavlanos, "On the convergence of a distributed Augmented Lagrangian method for nonconvex optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 9, pp. 4405–4420, 2017.

[31] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Springer, Berlin, Heidelberg, 1998.

[32] J. Nocedal and S. Wright, *Numerical optimization*. Springer New York, 2006.

[33] S. Frank and S. Rebennack, "An introduction to optimal power flow: Theory, formulation, and examples," *IIE transactions*, vol. 48, no. 12, pp. 1172–1197, 2016.

[34] A. Maffei, D. Meola, G. Marafioti, G. Palmieri, L. Iannelli, G. Mathisen, E. Bjerkan, and L. Glielmo, "Optimal power flow model with energy storage, an extension towards large integration of renewable energy sources." *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 9456–9461, 2014, 19th IFAC World Congress.

[35] A. Castillo, C. Laird, C. A. Silva-Monroy, J.-P. Watson, and R. P. O'Neill, "The unit commitment problem with ac optimal power flow constraints," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4853–4866, 2016.

[36] D. K. Molzahn, I. A. Hiskens *et al.*, "A survey of relaxations and approximations of the power flow equations," *Foundations and Trends® in Electric Energy Systems*, vol. 4, no. 1-2, pp. 1–221, 2019.

[37] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on power systems*, vol. 26, no. 1, pp. 12–19, 2010.

[38] ISO New England, "Hourly Real-Time System Demand," https://www.iso-ne.com/isoexpress/web/reports/load-and-demand/-/tree/dmnd-rt-hourly-sys, Online.

[39] I. Dunning, J. Huchette, and M. Lubin, "JuMP: A modeling language for mathematical optimization," *SIAM Review*, vol. 59, no. 2, pp. 295–320, 2017.

[40] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical programming*, vol. 106, no. 1, pp. 25–57, 2006.

[41] S. S. Vazhkudai *et al.*, "The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems," in *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2018, pp. 661–672.

[42] A. V. Fiacco, "Sensitivity analysis for nonlinear programming using penalty methods," *Mathematical programming*, vol. 10, no. 1, pp. 287–311, 1976.

[43] D. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

Anirudh Subramanyam is a postdoctoral researcher in the Mathematics and Computer Science Division at Argonne National Laboratory. He obtained his bachelor's degree from the Indian Institute of Technology, Bombay and his Ph.D. from Carnegie Mellon University, both in chemical engineering. His research interests are in computational methods for nonlinear and discrete optimization under uncertainty with applications in energy, transportation and process systems.

Youngdae Kim is a postdoctoral researcher in the Mathematics and Computer Science Division at Argonne National Laboratory. He received the B.Sc. and M.Sc. degrees in computer science and engineering from Pohang University of Science and Technology, Pohang, South Korea, and the Ph.D. degree in computer sciences from the University of Wisconsin-Madison. His research interests include distributed optimization methods using hardware accelerators with applications in power systems.

Michel Schanen is an assistant computer scientist in the Mathematics and Computer Science Division at Argonne National Laboratory. He received his M.Sc. in 2008 from Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, Germany and in 2014 his Ph.D. in computer science from RWTH Aachen, Germany. His interests lie in automatic differentiation and applications in large-scale computing.

François Pacaud is a postdoctoral researcher in the Mathematics and Computer Science Division at Argonne National Laboratory. He obtained his M.Sc. in 2015 from Mines ParisTech, Paris, France, and in 2018 his Ph.D. in applied mathematics from the École des Ponts ParisTech, Paris, France. His research interests encompasses stochastic and nonlinear optimization, with application in energy systems.

Mihai Anitescu is a senior computational mathematician in the Mathematics and Computer Science Division at Argonne National Laboratory and a professor in the Department of Statistics at the University of Chicago. He obtained his engineer diploma (electrical engineering) from the Polytechnic University of Bucharest in 1992 and his Ph.D. in applied mathematical and computational sciences from the University of Iowa in 1997. He specializes in the areas of numerical optimization, computational science, numerical analysis and uncertainty quantification in which he has published more than 100 papers in scholarly journals and book chapters. He has been recognized for his work in applied mathematics by his selection as a SIAM Fellow in 2019.