SIAM J. CONTROL OPTIM.
Vol. 60, No. 4, pp. 1970–1990

RUNNING PRIMAL-DUAL GRADIENT METHOD FOR TIME-VARYING NONCONVEX PROBLEMS*

YUJIE TANG†, EMILIANO DALL'ANESE‡, ANDREY BERNSTEIN§, AND STEVEN LOW¶

Abstract. This paper focuses on a time-varying constrained nonconvex optimization problem, and considers the synthesis and analysis of online regularized primal-dual gradient methods to track a Karush–Kuhn–Tucker (KKT) trajectory. The proposed regularized primal-dual gradient method is implemented in a running fashion, in the sense that the underlying optimization problem changes during the execution of the algorithms. In order to study its performance, we first derive its continuous-time limit as a system of differential inclusions. We then study sufficient conditions for tracking a KKT trajectory, and also derive asymptotic bounds for the tracking error (as a function of the time-variability of a KKT trajectory). Further, we provide a set of sufficient conditions for the KKT trajectories not to bifurcate or merge, and also investigate the optimal choice of the parameters of the algorithm. Illustrative numerical results for a time-varying nonconvex problem are provided.

Key words. time-varying optimization, primal-dual dynamics, tracking, nonconvex optimization, gradient methods, differential inclusion

AMS subject classifications. 34A60, 49M37, 90C30, 90C31

DOI. 10.1137/20M1371063

1. Introduction.

This paper focuses on continuous-time nonconvex optimization problems of the form

$$\min_{x\in\mathcal{X}(t)} \quad c(x,t) \qquad \text{s.t.} \quad f(x,t)\in\mathcal{K}, \tag{1.1}$$

where $t\in[0,S]$ is the temporal index for some $S>0$, $c:\mathbb{R}^n\times[0,S]\to\mathbb{R}$ and $f:\mathbb{R}^n\times[0,S]\to\mathbb{R}^m$ are twice continuously differentiable functions, $\mathcal{X}(t)$ is a closed convex subset of $\mathbb{R}^n$ for each t, and $\mathcal{K}\subseteq\mathbb{R}^m$ is a closed and convex cone. The optimization problem (1.1) can be associated with systems governed by possibly nonlinear physical or learning models, and can capture performance objectives and constraints that evolve over time. The interdisciplinary nature of time-varying optimization problems [37] is evident through a number of works in the areas of power systems [42, 24], communication systems [30, 12], transportation systems [44], robotic networks [17], and online learning [17, 32] just to mention a few. On the other hand, suitable modifications of (1.1) can model time-varying data processing problems under streaming of measurements such as matrix factorization [29] and sparse signal recovery [3].

A common approach in the existing literature is that, with a discretization $t=\tau\Delta_T$ for $\tau=0,1,\dots,T$ with sampling interval $\Delta_T=S/T$ [37], one can employ the following running (or online) primal-dual gradient algorithm to track the optimal

*Received by the editors October 19, 2020; accepted for publication (in revised form) March 1, 2022; published electronically July 7, 2022.
https://doi.org/10.1137/20M1371063
Funding: This work was supported by the ARPA-E through grant DE-AR0000699, and by the NSF through grants CCF 1637598, ECCS 1619352, CPS 1739355.

†School of Engineering and Applied Sciences, Harvard University, Allston, MA 02134 USA (yujietang@seas.harvard.edu).
‡Department of Electrical, Computer, and Energy Engineering, University of Colorado Boulder, Boulder, CO 80309 USA (emiliano.dallanese@colorado.edu).
§National Renewable Energy Laboratory, Golden, CO 80401 USA (andrey.bernstein@nrel.gov).
¶Departments of Computing and Mathematical Sciences and Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (slow@caltech.edu).

trajectory $(x^*(t), \lambda^*(t))$ at time instants $t = \tau\Delta_T$ for $\tau = 1, 2, \ldots, T$:

$$x_\tau = \mathcal{P}_{\mathcal{X}_\tau}\left[x_{\tau-1} - \alpha\left(\nabla c_\tau(x_{\tau-1}) + J_{f_\tau}(x_{\tau-1})^T\lambda_{\tau-1}\right)\right], \tag{1.2a}$$

$$\lambda_\tau = \mathcal{P}_{\mathcal{K}^\circ}\left[\lambda_{\tau-1} + \eta\alpha\left(f_\tau(x_{\tau-1}) - \epsilon\lambda_{\tau-1}\right)\right], \tag{1.2b}$$

where $c_\tau(\cdot) \coloneqq c(\cdot, \tau\Delta_T)$, $f_\tau(\cdot) \coloneqq f(\cdot, \tau\Delta_T)$, $\mathcal{X}_\tau \coloneqq \mathcal{X}(\tau\Delta_T)$, and $\mathcal{K}^\circ$ denotes the polar cone of $\mathcal{K}$; the parameters $\alpha > 0$ and $\eta > 0$ control the step sizes of the primal and dual updates; the parameter $\epsilon > 0$ controls regularization on the dual variable; and $\mathcal{P}_{\mathcal{X}}$ denotes the projection onto the convex set $\mathcal{X}$. We call this algorithm the *Running Regularized Primal-Dual gradient algorithm* (RReg-PD). RReg-PD can be viewed as an extension of the regularized primal-dual algorithm in [27] to the time-varying situation, and its performance for convex problems has been partially analyzed in, e.g., [38, 5].

On the other hand, in many real-world applications such as real-time optimal power flow [42], wind farm operation [31], transportation systems [6], and data processing [29] the objective function $c(x, t)$ and/or the constraint function $f(x, t)$ may not be (globally) convex in x for each t. While studies have shown that RReg-PD may still work well in practice, its theoretical analysis is still lacking. Moreover, while [38] shows that the tracking error of RReg-PD will always be bounded provided that $c(x, t)$ is uniformly strongly convex in x for all t and that the step sizes are sufficiently small, how to choose the algorithmic parameters to minimize the tracking error to improve the tracking performance is still unclear. These issues motivate our work on theoretical analysis of RReg-PD for possibly nonconvex problems.

Contributions. This paper extends [41] and conducts a comprehensive theoretical analysis on the performance of RReg-PD for nonconvex smooth problems from the perspective of its continuous-time limit. Specific contributions include the following.

1. We consider the limiting case where the algorithm can afford an update in infinitesimal time, i.e., the continuous-time limit of RReg-PD (C-RReg-PD). We show that the continuous-time counterpart of the discrete-time algorithm is given by a system of differential inclusions that has been studied in the literature under the name of *perturbed sweeping processes* [11, 10]; the discrete-time algorithm was referred to as the *catching algorithm* of the perturbed sweeping processes.
2. We provide sufficient conditions for guaranteeing bounded tracking error of C-RReg-PD in the situation where the objective $c(x, t)$ and the constraint function $f(x, t)$ are smooth but can be nonconvex. Existing works on regularized primal-dual methods for convex problems [27, 38, 48] only proved that gradient-based iterative methods approach an approximate Karush–Kuhn–Tucker (KKT) point; instead, we provide analytical results in terms of tracking a KKT trajectory (as opposed to an approximate KKT trajectory), and our results are applicable to convex and nonconvex problems.
3. We also provide sufficient conditions under which the KKT points for a given time instant will always be isolated; that is, bifurcations or merging of KKT trajectories do not happen. We shall also show that these conditions for isolated KKT trajectories are implied by the sufficient conditions for bounded tracking error of C-RReg-PD.
4. Finally, we provide qualitative results and discussions on how the algorithmic parameters can be chosen to achieve good tracking performance in practice.

1.1. Related work.

Time-varying optimization. Reference [33] is one of the early papers that consider a time-varying optimization setting similar to this paper, which derived a tracking error bound of the running gradient descent algorithm for unconstrained convex problems. Recent years have witnessed considerable advances in time-varying optimization and online learning. A sample of works include [38, 5], which proposed and analyzed an online double smoothing method, and [36], which proposed a unified framework for time-varying convex optimization using averaged operators. Time-varying fixed-point methods were discussed in [4] for contraction maps with asynchronous updates. Time-varying projected gradient dynamics were considered in [24], where the analysis leveraged the notion of temporal tangent cones. Recently, [16] showed that the inherent temporal variation of a time-varying optimization problem could reshape the landscape and result in escaping spurious local minimum; this work was extended to nonconvex problems with linear equality constraints in [20]. For convex settings, we refer the reader to additional references collected in the recent tutorial [37].

Dynamic regret of online optimization. There has been abundant research on the theory of dynamic regret for online optimization [25, 32, 21, 28]. The dynamic regret is defined as the accumulated losses compared with the sequence of optimal decisions, and researchers are interested in the asymptotic growth rate of the dynamic regret. For example, in [32], the authors considered online convex optimization with uniformly strongly convex and smooth objectives, and showed $O(C_T)$ bound for the dynamic regret, where C_T is the path variation. Reference [21] considered online optimization with weakly pseudo-convex objectives, and showed $O(\sqrt{T + V_T T})$ bound for the dynamic regret where V_T is the loss variation. The paper [28] proposed a second-order method for online nonconvex optimization. We note that most works on online optimization seem to focus on situations with sublinear regret, which is different from our situation where a linear growth rate can occur and the big-O asymptotic growth rate is not enough to characterize the performance.

Parametric optimization. Time-varying optimization is closely related to parametric optimization [23, 47, 18], and path-following algorithms have been proposed to find a sequence of points that approximates the optimal trajectory of (1.1). On the other hand, path-following algorithms in general incorporate second-order information so that the generated sequence will converge to the optimal trajectory as the sampling interval shrinks to zero, and may also incorporate temporal derivatives to accelerate the convergence; this is different from the algorithm considered in this paper which only employs first-order spatial derivatives.

Gradient-based methods for nonconvex problems. Complexity of gradient-based methods converging to stationary points for nonconvex problems have been studied in both the centralized and distributed settings [9, 40]. In addition, some recent works show that adding appropriate random perturbations in gradient-based methods enables escaping from saddle points [26, 2]; these works only treat static problems, and extensions to time-varying settings are still open problems and worth investigation.

Continuous-time gradient flows. The idea of studying discrete-time algorithms by resorting to their continuous-time limits has a long history, and it is well known that the gradient flow $\dot{x} = -\nabla f(x)$ characterizes the limiting behavior of the gradient descent method. For more sophisticated settings, [13, 43] analyzed the convergence of the saddle point dynamics, [22, 34] considered continuous-time dynamics for distributed optimization problems, [14] studied primal-dual dynamics for distributed optimization with time-varying components via contraction analysis, and [39, 35] studied accelerated gradient descent methods from the perspective of discretization of differential

equations.

Notation. For a twice differentiable real-valued function f, its Hessian at x will be denoted by $\nabla^2 f(x)$. For a function $f(x,t)$ that is continuously differentiable in x, its gradient with respect to x at (u,s) will be denoted by $\nabla_x f(u,s)$. If $f(x,s)$ is twice differentiable with respect to x, its Hessian with respect to x at (u,s) will be denoted by $\nabla^2_{xx} f(u,s)$. For a differentiable vector-valued function f, its Jacobian matrix evaluated at x will be denoted by $J_f(x)$. For vector-valued $f(x,t)$ that is differentiable in x, its Jacobian matrix with respect to x at (u,s) is denoted by $J_{f,x}(u,s)$.

Given a convex set C, its normal cone at $x \in C$, defined by $\{y : y^T(z-x) \leq 0 \ \forall z \in C\}$, is denoted by $N_C(x)$. The projection onto C will be denoted by $\mathcal{P}_C$. For a convex cone C, we denote its polar cone by $C^\circ = \{y : y^T z \leq 0 \ \forall z \in C\}$. We denote $\mathbb{R}_+ = [0, +\infty)$ and $\mathbb{R}_{++} = (0, +\infty)$. The identity matrix is denoted by I, or $I_n \in \mathbb{R}^{n \times n}$ when the dimensions need to be specified. We denote the ℓ_2-norm by $\|\cdot\|$.

2. Problem formulation. As has been introduced in section 1, this paper considers the following time-varying optimization problem:

$$\min_{x \in \mathcal{X}(t)} \quad c(x,t) \qquad \text{s.t.} \quad f(x,t) \in \mathcal{K}. \tag{1.1}$$

Notice that the temporal index t varies in a continuous interval $[0, S]$. We shall see that this setting allows us to analyze the C-RReg-PD:

$$x_\tau = \mathcal{P}_{\mathcal{X}_\tau}\left[x_{\tau-1} - \alpha\left(\nabla c_\tau(x_{\tau-1}) + J_{f_\tau}(x_{\tau-1})^T \lambda_{\tau-1}\right)\right], \tag{1.2a}$$

$$\lambda_\tau = \mathcal{P}_{\mathcal{K}^\circ}\left[\lambda_{\tau-1} + \eta\alpha\left(f_\tau(x_{\tau-1}) - \epsilon\lambda_{\tau-1}\right)\right]. \tag{1.2b}$$

We remind the readers of the notation $c_\tau(\cdot) \coloneqq c(\cdot, \tau\Delta_T)$, $f_\tau(\cdot) \coloneqq f(\cdot, \tau\Delta_T)$, $\mathcal{X}_\tau \coloneqq \mathcal{X}(\tau\Delta_T)$, and that $\mathcal{K}^\circ$ is the polar cone of $\mathcal{K}$; also recall that $\alpha > 0$ and $\eta > 0$ control the step sizes of the primal and dual updates. RReg-PD can be derived by applying one single iteration of the vanilla primal-dual gradient algorithm on a "regularized" Lagrangian [38], and the parameter $\epsilon > 0$ stems from a strongly concave regularization term on the dual variable λ.

For the purpose of theoretical analysis, we decompose the function $f(x,t)$ into a convex part $f^c(x,t)$ and a nonconvex part $f^{nc}(\cdot,t)$ as

$$f(x,t) \coloneqq f^c(x,t) + f^{nc}(x,t), \tag{2.1}$$

where for each t, the function $f^c(\cdot,t)$ is *convex with respect to* $-\mathcal{K}$ in the sense that the set $\{f^c(x,t) + y : -y \in \mathcal{K}\}$ is convex.[1] Note that if we let $\mathcal{K} = (-\mathbb{R}^p_+) \times \{0\}^{m-p}$, then $f(x,t) \in \mathcal{K}$ corresponds to p inequality constraints and $m-p$ equality constraints, and the components of $f^c(x,t)$ comprise p convex functions and $m-p$ affine functions.

We impose the following assumptions regarding the problem (1.1).

Assumption 2.1.

1. For each $t \in [0,S]$, the set $\mathcal{X}(t)$ is compact and convex. Moreover, $\mathcal{X} : [0,S] \to 2^{\mathbb{R}^n}$ is a κ_1-Lipschitz set-valued map for some $\kappa_1 > 0$, i.e.,

$$d_H(\mathcal{X}(t_1), \mathcal{X}(t_2)) \leq \kappa_1 |t_1 - t_2| \quad \forall t_1, t_2 \in [0,S],$$

where d_H denotes the Hausdorff distance.

[1] Equivalently, we have that for any $\lambda \in \mathcal{K}^\circ$, the function $\lambda^T f^c(\cdot,t)$ is convex in the ordinary sense; see [7, section 2.3.5] for more details.

2. For each $t \in [0, S]$, the functions $c(x,t)$, $f^c(x,t)$, and $f^{nc}(x,t)$ are twice continuously differentiable over $x \in \mathbb{R}^n$. In addition, $\nabla^2_{xx} f_i^c(x,t), \nabla^2_{xx} f_i^{nc}(x,t)$ for $i = 1, \ldots, m$ and $\nabla^2_{xx} c(x,t)$ are continuous over $(x,t) \in \mathbb{R}^n \times [0,S]$.
3. There exists a Lipschitz continuous trajectory $z^*(t) = \big(x^*(t), \lambda^*(t)\big)$, $t \in [0,S]$, such that for each $t \in [0,S]$,

$$(x^*(t), \lambda^*(t)) \in \mathcal{X}(t) \times \mathcal{K}^\circ, \tag{2.2a}$$

$$\nabla_x c(x^*(t),t) + J_{f,x}(x^*(t),t)^T \lambda^*(t) \in -N_{\mathcal{X}(t)}(x^*(t)), \tag{2.2b}$$

$$\lambda^*(t) \in N_{\mathcal{K}}(f(x^*(t),t)). \tag{2.2c}$$

The conditions (2.2) are just the KKT conditions of (1.1) for each $t \in [0,S]$. Indeed, [19, Theorem 2A.9] shows that if we assume the following constraint qualification condition for every $t \in [0,S]$:

$$\begin{aligned} &\text{there is no } \lambda \in \mathcal{K}^\circ \backslash \{0\} \text{ such that} \\ &\lambda^T f(x^*(t),t) = 0 \text{ and } -J_{f,x}(x^*(t),t)^T \lambda \in N_{\mathcal{X}(t)}(x^*(t)), \end{aligned} \tag{2.3}$$

where $x^*(t)$ is a local optimal solution to (1.1), then there exists an associated optimal Lagrange multiplier $\lambda^*(t)$ that satisfies the KKT conditions (2.2) for each $t \in [0,S]$.

Remark 2.2. In general, there can be multiple KKT points of (1.1) that move in $\mathbb{R}^n \times \mathbb{R}^m_+$ as time proceeds, which form multiple trajectories that can appear, terminate, bifurcate, or merge during $(0,S)$. Reference [23] presents a comprehensive theory of the structures and singularities of KKT trajectories. Reference [18] shows that strong regularity for generalized equations is a key concept for establishing the existence of Lipschitz continuous KKT trajectories. Here, we assume the existence of a Lipschitz continuous trajectory of KKT points over $t \in [0,S]$, and arbitrarily select one of them when multiple Lipschitz continuous trajectories exist. We denote this trajectory by $z^*(t)$, and focus on this trajectory in most part of our study. How to deal with KKT trajectories with discontinuities will be left for future investigations.

3. Continuous-time limit of RReg-PD. In this section, we formulate the continuous-time limit of the discrete-time RReg-PD algorithm. The hope is that, by resorting to the continuous-time limit, we can obtain some new insights on RReg-PD.

We first present a precise description of what we mean by "continuous-time limit." Let $z_0 = (x_0, \lambda_0) \in \mathcal{X}(0) \times \mathcal{K}^\circ$ be some fixed initial point. We note that the discrete-time RReg-PD (1.2) can be abstractly written as

$$z_\tau^{(T)} = \mathcal{P}_{\mathcal{C}(\tau\Delta_T)} \left[z_{\tau-1}^{(T)} + \alpha_T \cdot \Phi\big(z_{\tau-1}^{(T)}, \tau\Delta_T\big) \right], \quad \tau = 1, \ldots, T, \qquad z_0^{(T)} = z_0, \tag{3.1}$$

where

$$\Phi\left(\begin{bmatrix} x \\ \lambda \end{bmatrix}, t\right) \coloneqq \begin{bmatrix} -\nabla c(x,t) - J_{f,x}(x,t)^T \lambda \\ \eta(f(x,t) - \epsilon\lambda) \end{bmatrix}, \qquad \mathcal{C}(t) \coloneqq \mathcal{X}(t) \times \mathcal{K}^\circ$$

for $x \in \mathbb{R}^n, \lambda \in \mathbb{R}^m$, and $t \in [0,S]$. Note that the step size α_T can be dependent on T (but we fix $\eta > 0$). We also add the superscript (T) to the iterates $z_\tau^{(T)}$ so that the sequences of iterates generated using different sampling intervals Δ_T can be distinguished. We then consider the linear interpolation of $\big(z_\tau^{(T)}\big)_{\tau=1,\ldots,T}$ defined by

$$z^{[T]}(t) = \frac{\tau\Delta_T - t}{\Delta_T} z_{\tau-1}^{[T]} + \frac{t - (\tau-1)\Delta_T}{\Delta_T} z_\tau^{[T]}, \qquad t \in [(\tau-1)\Delta_T, \tau\Delta_T] \tag{3.2}$$

for each $\tau = 1, 2, \ldots, T$. We see that for each $T \in \mathbb{N}$, there is a corresponding continuous trajectory $z^{[T]} : [0, S] \to \mathbb{R}^{n+m}$ that linearly connects the iterates of (3.1). We say that a function $z : [0, S] \to \mathbb{R}^{n+m}$ is a continuous-time limit of (3.1) with the initial point z_0, if $z^{[T]}$ has a subsequence $z^{[T_k]}$ such that

$$\lim_{k\to\infty} z^{[T_k]}(t) = z(t) \qquad \forall t \in [0, S], \tag{3.3}$$

i.e., $z(t)$ can be viewed as a limiting trajectory when we take Δ_T to be infinitesimal.

It turns out that the continuous-time limit of discrete-time iterations in the form of (3.1) has been studied in the literature [10, 11]. Specifically, we have the following theorem, whose proof is given in Appendix A.

THEOREM 3.1. *Let $\beta > 0$ be arbitrary and fixed. Suppose Assumption* 2.1 *and the following technical conditions hold:*

(i) *The gradient $\nabla_x c(x, t)$ is continuous over $(x, t) \in \bigcup_{t\in[0,S]} \mathcal{X}(t) \times [0, S]$, and there exists $\kappa_2 > 0$ such that*

$$\|\nabla_x c(x, t)\| \le \kappa_2(1 + \|x\|) \qquad \forall (x, t) \in \bigcup\nolimits_{t\in[0,S]} \mathcal{X}(t) \times [0, S]. \tag{3.4}$$

(ii) *The function $f(x, t)$ is continuous, and the Jacobian $J_{f,x}(x, t)$ is bounded and continuous over $(x, t) \in \bigcup_{t\in[0,S]} \mathcal{X}(t)\times[0, S]$. Moreover, there exists some integrable function $l : [0, S] \to [0, +\infty)$ such that for each $t \in [0, S]$,*

$$\sup_{x\in\mathcal{X}(t)} \left\|\nabla^2_{xx} c(x, t)\right\| \le l(t), \qquad \sup_{x\in\mathcal{X}(t)} \left\|\nabla^2_{xx} f_j(x, t)\right\| \le l(t) \qquad \forall j = 1, \ldots, m.$$

Then given any initial point $(x(0), \lambda(0)) = (x_0, \lambda_0) \in \mathcal{X}(0)\times\mathcal{K}^\circ$, we have the following:

1. *There exists an absolutely continuous solution to the differential inclusions*

$$-\frac{d}{dt}x(t) - \beta\left(\nabla_x c(x(t), t) + J_{f,x}(x(t), t)^T\lambda(t)\right) \in N_{\mathcal{X}(t)}(x(t)), \tag{3.5a}$$

$$-\frac{d}{dt}\lambda(t) + \eta\beta\left(f(x(t), t) - \epsilon\lambda(t)\right) \in N_{\mathcal{K}^\circ}\big(\lambda(t)\big), \tag{3.5b}$$

   and this absolutely continuous solution is also Lipschitz continuous.

2. *A continuous-time limit of* (3.1) *with the initial point z_0 is given by a Lipschitz continuous solution to* (3.5)*, provided that we set the step sizes as $\alpha_T = \beta\Delta_T$. In addition, the convergence in* (3.3) *is uniform.*

Note that (3.5) can also be compactly written as

$$-\frac{d}{dt}z(t) + \beta\Phi(z(t), t) \in N_{\mathcal{C}(t)}(z(t)). \tag{3.6}$$

This type of differential inclusions is called *perturbed sweeping processes* [10, 11, 1], and the discrete-time algorithm (3.1) with $\alpha_T = \beta\Delta_T$ is called the *catching algorithm* associated with (3.6).

Remark 3.2. We point out that when the set $\mathcal{C}(t)$ is time-varying, the perturbed sweeping process (3.6) is, in general, not equivalent to the projected dynamical systems

$$\frac{dz(t)}{dt} = \mathcal{P}_{T_{\mathcal{C}(t)}(z(t))}\left[\beta\Phi(z(t), t)\right] \quad \text{or} \quad \frac{dz(t)}{dt} = \lim_{s\to 0^+} \frac{\mathcal{P}_{\mathcal{C}(t)}\left[z(t)+s\beta\Phi(z(t), t)\right] - z(t)}{s}. \tag{3.7}$$

In fact the projected dynamical systems in (3.7) may not even have a solution. A simple example is $\mathcal{C}(t) = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \ge t\}$ with $\Phi(z, t) = 0$. Under the initial

condition $(x_1(0), x_2(0)) = (0, 0)$, (3.6) admits the solution $x_1(t) = t, x_2(t) = 0$, but neither formulation in (3.7) has a solution. This is different from the case where $\mathcal{C}(t)$ is static [8]. The work [24] introduced a formulation similar to (3.7) based on the notion of temporal tangent cones that generalizes tangent cones to time-varying settings.

We shall also refer to the differential inclusions (3.5) as the continuous-time limit of the discrete-time RReg-PD (3.1), and call it C-RReg-PD for short.

4. Tracking performance. After formulating the C-RReg-PD, we proceed to conduct theoretical analysis on its performance.

We first clarify how we measure the performance of C-RReg-PD. Let $z(t), t \in [0, S]$, be a trajectory that satisfies the differential inclusions (3.5). We define the *tracking error* with respect to a KKT trajectory $z^*(t) = (x^*(t), \lambda^*(t))$ at time $t \in [0, S]$ by

$$\|z(t) - z^*(t)\|_\eta = \sqrt{\|x(t) - x^*(t)\|^2 + \eta^{-1}\|\lambda(t) - \lambda^*(t)\|^2},$$

where we also introduce the norm $\|\cdot\|_\eta$ by $\|z\|_\eta := \sqrt{\|x\|^2 + \eta^{-1}\|\lambda\|^2}$ for $z = (x, \lambda) \in \mathbb{R}^{n+m}$. The tracking error quantifies the suboptimality of $z(t)$ in terms of the distance to the KKT point $z^*(t)$. We are interested in the factors that affect the tracking error, especially the conditions under which a bounded tracking error can be guaranteed, and how to reduce the tracking error by properly choosing the algorithmic parameters.

Intuitively, it's natural to think that the faster the KKT point moves, the more difficult it is for C-RReg-PD to track the KKT point, which leads to larger tracking errors. We therefore introduce the quantity

$$\sigma_\eta := \sup\left\{\frac{\|z^*(t_2) - z^*(t_1)\|_\eta}{|t_2 - t_1|} : t_1, t_2 \in [0, S], t_1 \neq t_2\right\},$$

which can be considered as an upper bound on the "speed" of the KKT point.

We then introduce some auxiliary quantities for further characterizing (1) the local convexity and (2) the local nonlinearity of the time-varying problem (1.1). By "local" we mean that these quantities are determined by the behavior of $c(x, t)$ and $f(x, t)$ within a neighborhood of some specified radius $\delta > 0$ around the KKT point $z^*(t)$.

1. $\Lambda_m(\delta)$ *for local convexity*: Let $\mathcal{L}^{nc}(x, \lambda, t) := c(x, t) + \lambda^T f^{nc}(x, t)$ denote the "nonconvex part" of the Lagrangian for $t \in [0, S]$. We define

$$\overline{H}_{\mathcal{L}^{nc}}(u, t) := \int_0^1 \nabla^2_{xx}\mathcal{L}^{nc}(x^*(t) + \theta u, \lambda^*(t), t)\, d\theta, \tag{4.1}$$

$$\overline{H}_{f_i^c}(u, t) := \int_0^1 2(1-\theta)\nabla^2_{xx} f_i^c(x^*(t) + \theta u, t)\, d\theta, \qquad i = 1, \ldots, m. \tag{4.2}$$

The matrix $\overline{H}_{\mathcal{L}^{nc}}(u, t)$ can be interpreted as the averaged Hessian of $\mathcal{L}^{nc}$ around $x^*(t)$ along the vector u, and similarly $\overline{H}_{f_i^c}(u, t)$ can be viewed as the (weighted) averaged Hessian matrix of the convex part of the ith constraint along the vector u.

We now introduce

$$\Lambda_m(\delta) := \inf_{t\in[0,S]} \inf_{\substack{u:\|u\|\le\delta,\\ x^*(t)+u\in\mathcal{X}(t)}} \lambda_{\min}\Big(\overline{H}_{\mathcal{L}^{nc}}(u, t) + \frac{1}{2}\sum\nolimits_{i=1}^m \lambda_i^*(t)\overline{H}_{f_i^c}(u, t)\Big). \tag{4.3}$$

Roughly speaking, $\Lambda_m(\delta)$ characterizes how convex the problem (1.1) is in the neighborhood of the optimal point of radius δ. In the special case where $c(\cdot,t)$ is μ-strongly convex for all t for some $\mu>0$, we have $\Lambda_m(\delta)\geq\mu$.

2. $M_{nc}(\delta)$ *for local nonlinearity*: This quantity is defined by

$$M_{nc}(\delta) \coloneqq \sup_{t\in[0,S]} \sup_{\substack{u:\|u\|\leq\delta,\\ x^*(t)+u\in\mathcal{X}(t)}} \sqrt{\sum\nolimits_{i=1}^{n} \lambda_{\max}^2\big(\overline{H}_{f_i^{nc}}(u,t)\big)}. \tag{4.4}$$

Here the matrix $\overline{H}_{f_i^{nc}}(u,t)$ is defined by

$$\overline{H}_{f_i^{nc}}(u,t) \coloneqq \int_0^1 2\theta\,\nabla_{xx}^2 f_i^{nc}(x^*(t)+\theta u,t)\,d\theta, \qquad i=1,\dots,m.$$

We point out that the quantity $M_{nc}(\delta)$ characterizes the local nonlinearity of the *nonconvex* part of the constraints in the neighborhood of $x^*(t)$ of radius δ.

Finally, we denote

$$M_\lambda \coloneqq \sup_{t\in[0,S]} \|\lambda^*(t)\|,$$

i.e., the maximum magnitude of the optimal dual variable during the period $[0,S]$.

We are now ready to present our results on the tracking performance.

THEOREM 4.1. *Suppose there exist $\delta>0$, $\beta>0$, $\eta>0$, and $\epsilon>0$ such that*

$$\beta^{-1}\sigma_\eta < \delta\,\gamma(\delta,\eta,\epsilon) - \sqrt{\eta}\epsilon M_\lambda, \tag{4.5}$$

where

$$\gamma(\delta,\eta,\epsilon) \coloneqq \min\{\Lambda_m(\delta),\eta\epsilon\} - \frac{\sqrt{\eta}}{4}\delta M_{nc}(\delta). \tag{4.6}$$

Let $z(t)$ be a Lipschitz continuous solution to (3.5) *with $\|z(0)-z^*(0)\|_\eta<\delta$. Then for all $t\in[0,S]$,*

$$\begin{aligned}\|z(t)-z^*(t)\|_\eta < {}& \frac{\beta^{-1}\sigma_\eta+\sqrt{\eta}\epsilon M_\lambda}{\gamma(\delta,\eta,\epsilon)} \\ & + e^{-\beta\gamma(\delta,\eta,\epsilon)\,t}\left(\|z(0)-z^*(0)\|_\eta - \frac{\beta^{-1}\sigma_\eta+\sqrt{\eta}\epsilon M_\lambda}{\gamma(\delta,\eta,\epsilon)}\right).\end{aligned} \tag{4.7}$$

Proof. For notational simplicity, we shall frequently use $z=(x,\lambda)$ to denote the trajectory of C-RReg-PD and $z^*=(x^*,\lambda^*)$ to denote the KKT trajectory, where we suppress their dependence on t.

To begin our proof, we notice that the first differential inclusion (3.5a) implies

$$0\geq(x^*-x)^T\left(-\dot{x}-\beta\nabla_x\mathcal{L}^{nc}(x,\lambda^*,t)-\beta J_{f^c,x}(x,t)^T\lambda+\beta J_{f^{nc},x}(x,t)^T\big(\lambda^*-\lambda\big)\right),$$

and by the KKT condition (2.2b),

$$0\geq(x-x^*)^T\left(-\beta\nabla_x\mathcal{L}^{nc}(x^*,\lambda^*,t)-\beta J_{f^c,x}(x^*,t)^T\lambda^*\right).$$

By multiplying the previous two inequalities by -1, summing them up and adding $(x-x^*)^T\dot{x}$ on both sides, we get

$$\begin{aligned}(x-x^*)^T\dot{x} \le & -\beta(x-x^*)^T\big[\nabla_x\mathcal{L}^{nc}(x,\lambda^*,t)-\nabla_x\mathcal{L}^{nc}(x^*,\lambda^*,t)\\ & +J_{f^c,x}(x,t)^T\lambda-J_{f^c,x}(x^*,t)^T\lambda^*+J_{f^{nc},x}(x,t)^T\big(\lambda-\lambda^*\big)\big]\\ = & -\beta(x-x^*)^T\big[B_{\mathcal{L}^{nc}}(t)\,(x-x^*)+J_{f^{nc},x}(x,t)^T\big(\lambda-\lambda^*\big)\\ & +J_{f^c,x}(x,t)^T\lambda-J_{f^c,x}(x^*,t)^T\lambda^*\big],\end{aligned}$$

where we denote $B_{\mathcal{L}^{nc}}(t)\coloneqq\overline{H}_{\mathcal{L}^{nc}}(x-x^*,t)$. Next, by the second differential inclusion (3.5b),

$$0\ge(\lambda^*-\lambda)^T\big(-\dot{\lambda}+\eta\beta\,(f(x,t)-\epsilon\lambda)\big),$$

and by the KKT condition (2.2c), we see that $0\ge\eta\beta\big(\lambda-\lambda^*\big)^T f(x^*,t)$. Similarly, by multiplying these two inequalities by -1, summing them up, and adding $(\lambda-\lambda^*)^T\dot{\lambda}$ on both sides, we get

$$\big(\lambda-\lambda^*\big)^T\dot{\lambda}\le\eta\beta\big(\lambda-\lambda^*\big)^T\big(f(x,t)-f(x^*,t)\big)-\eta\beta\epsilon\big\|\lambda-\lambda^*\big\|^2-\eta\beta\epsilon\big(\lambda-\lambda^*\big)^T\lambda^*.$$

By taking the time derivative of $\frac{1}{2}\|z-z^*\|_\eta^2$ and using the previous bounds, we get

(4.8)
$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|z-z^*\|_\eta^2=&(x-x^*)^T\dot{x}+\eta^{-1}\big(\lambda-\lambda^*\big)^T\dot{\lambda}-(z-z^*)^T\begin{bmatrix}I_n & \\ & \eta^{-1}I_m\end{bmatrix}\dot{z}^*\\ \le & -\beta(x-x^*)^TB_{\mathcal{L}^{nc}}(t)(x-x^*)-\beta\epsilon\|\lambda-\lambda^*\|^2\\ & -\beta(x-x^*)^T\left(J_{f^c,x}(x,t)^T\lambda(t)-J_{f^c,x}(x^*,t)^T\lambda^*\right)\\ & -\beta(x-x^*)^TJ_{f^{nc},x}(x,t)^T\big(\lambda-\lambda^*\big)+\beta\left(f(x,t)-f(x^*,t)\right)^T\left(\lambda-\lambda^*\right)\\ & -\beta\epsilon{\lambda^*}^T\big(\lambda-\lambda^*\big)+\sigma_\eta\|z-z^*\|_\eta.\end{aligned}$$

Note that for the third to the fifth terms on the the right-hand side of (4.8), we have

$$\begin{aligned}& -(x-x^*)^T\left(J_{f^c,x}(x,t)^T\lambda-J_{f^c,x}(x^*,t)^T\lambda^*\right)\\ & -(x-x^*)^TJ_{f^{nc},x}(x,t)^T\big(\lambda-\lambda^*\big)+\left(f(x,t)-f(x^*,t)\right)^T\left(\lambda-\lambda^*\right)\\ = & -\lambda^T\left(f^c(x^*,t)-f(x,t)-J_{f^c,x}(x,t)(x^*-x)\right)\\ & -{\lambda^*}^T\left(f^c(x,t)-f(x^*,t)-J_{f^c,x}(x^*,t)(x-x^*)\right)\\ & -\big(\lambda-\lambda^*\big)^T\left(f^{nc}(x^*,t)-f^{nc}(x,t)-J_{f^{nc},x}(x,t)(x^*-x)\right).\end{aligned}$$

Furthermore, the convexity of f^c with respect to $-\mathcal{K}$ and the fact $\lambda\in\mathcal{K}^\circ$ imply that $\lambda^Tf^c(\cdot,t)$ is a convex function (in the ordinary sense), which further leads to $\lambda^T(f^c(x^*,t)-f(x,t)-J_{f^c,x}(x,t)(x^*-x))\ge0$. In addition,

$${\lambda^*}^T(f^c(x,t)-f(x^*,t)-J_{f^c,x}(x^*,t)(x-x^*))=(x-x^*)^T\Big(\frac{1}{2}\sum\nolimits_{i=1}^m\lambda_i^*B_{f_i^c}(t)\Big)(x-x^*),$$

where we denote $B_{f_i^c}(t)\coloneqq\overline{H}_{f_i^c}(x-x^*,t)$. Plugging them into (4.8) leads to

(4.9)
$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|z-z^*\|_\eta^2\le & -\beta(x-x^*)^T\Big(B_{\mathcal{L}^{nc}}(t)+\frac{1}{2}\sum\nolimits_{i=1}^m\lambda_i^*B_{f_i^c}(t)\Big)(x-x^*)-\beta\epsilon\big\|\lambda-\lambda^*\big\|^2\\ & -\beta\big(\lambda-\lambda^*\big)^T\big(f^{nc}(x^*,t)-f^{nc}(x,t)-J_{f^{nc},x}(x,t)(x^*-x)\big)\\ & -\beta\epsilon{\lambda^*}^T\big(\lambda-\lambda^*\big)+\sigma_\eta\|z-z^*\|_\eta.\end{aligned}$$

To further analyze the right-hand side of (4.9) and bound $\|z - z^*\|_\eta$, we need to partition the time domain $[0, S]$ into small intervals. Let κ_1 be the Lipschitz constant of $z(t)$ with respect to the norm $\|\cdot\|_\eta$, and define

$$\tilde{\Delta} := \frac{1}{2(\kappa_1 + \sigma_\eta)} \left(\delta - \max \left\{ \|z(0) - z^*(0)\|_\eta , \frac{\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda}{\gamma(\delta, \eta, \epsilon)} \right\} \right).$$

We note that $\tilde{\Delta} > 0$ under the given conditions. We shall prove by induction that

$$\|z(t) - z^*(t)\|_\eta \le \max \left\{ \|z(0) - z^*(0)\|_\eta , \frac{\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda}{\gamma(\delta, \eta, \epsilon)} \right\} \tag{4.10}$$

for $t \in \left[(k-1)\tilde{\Delta}, k\tilde{\Delta}\right] \cap [0, S]$ for each $k = 1, \ldots, \lceil S/\tilde{\Delta} \rceil$.

Obviously, (4.10) holds for $t = 0$. Now assume that (4.10) holds for $t = k\tilde{\Delta}$ for some k. We then have

$$\begin{aligned} \|z(t) - z^*(t)\|_\eta &\le \left\|z(t) - z\left(k\tilde{\Delta}\right)\right\|_\eta + \left\|z\left(k\tilde{\Delta}\right) - z^*\left(k\tilde{\Delta}\right)\right\|_\eta + \left\|z^*\left(k\tilde{\Delta}\right) - z^*(t)\right\|_\eta \\ &\le (\kappa_1 + \sigma_\eta)\tilde{\Delta} + \left\|z(k\tilde{\Delta}) - z^*(k\tilde{\Delta})\right\|_\eta < \delta \end{aligned}$$

for any $t \in \left[k\tilde{\Delta}, (k+1)\tilde{\Delta}\right] \cap [0, S]$, where the second inequality follows by the definitions of κ_1 and σ_η. Thus, we can use the definition of $M_{nc}(\delta)$ and the mean value theorem of calculus to get

$$\|f^{nc}(x,t) + J_{f^{nc},x}(x,t)(x^* - x) - f^{nc}(x^*, t)\| \le \frac{M_{nc}(\delta)}{2} \|x - x^*\|^2 \tag{4.11}$$

for $t \in \left[k\tilde{\Delta}, (k+1)\tilde{\Delta}\right] \cap [0, S]$. Moreover, by Young's inequality,

$$\|x - x^*\| \left\|\lambda - \lambda^*\right\| \le \frac{1}{2} \left(\sqrt{\eta} \|x - x^*\|^2 + \sqrt{\eta}^{-1} \left\|\lambda - \lambda^*\right\|^2 \right) = \frac{1}{2}\sqrt{\eta} \|z - z^*\|_\eta^2 . \tag{4.12}$$

Combining (4.11) and (4.12) with (4.9), we get, for $t \in \left[k\tilde{\Delta}, (k+1)\tilde{\Delta}\right] \cap [0, S]$,

(4.13)

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\|z - z^*\|_\eta^2 \\ &\le -\beta(x - x^*)^T \Big(B_{\mathcal{L}^{nc}}(t) + \frac{1}{2}\sum\nolimits_{i=1}^m \lambda_i^* B_{f_i^c}(t) \Big)(x - x^*) - \beta\epsilon \left\|\lambda - \lambda^*\right\|^2 \\ &\quad + \frac{\beta M_{nc}(\delta)}{2} \|x - x^*\|^2 \left\|\lambda - \lambda^*\right\| - \beta\epsilon {\lambda^*}^T \left(\lambda - \lambda^*\right) + \sigma_\eta \|z - z^*\|_\eta \\ &\le -\beta \left(\min\{\Lambda_m(\delta), \eta\epsilon\} - \frac{\sqrt{\eta}}{4}\delta M_{nc}(\delta) \right) \|z - z^*\|_\eta^2 + \beta \left(\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda \right) \|z - z^*\|_\eta, \end{aligned}$$

where the definitions of $\Lambda_m(\delta)$ and M_λ and $\|x - x^*\| \le \delta$ are employed in the second step. By the condition (4.5), we can apply Corollary B.2 (see Appendix B) and get

$$\begin{aligned} \|z(t) - z^*(t)\|_\eta \le\ & e^{-\beta\gamma(\delta,\eta,\epsilon)(t - k\tilde{\Delta})} \left(\left\|z\left(k\tilde{\Delta}\right) - z^*\left(k\tilde{\Delta}\right)\right\|_\eta - \frac{\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda}{\gamma(\delta, \eta, \epsilon)} \right) \\ &+ \frac{\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda}{\gamma(\delta, \eta, \epsilon)}, \qquad t \in \left[k\tilde{\Delta}, (k+1)\tilde{\Delta}\right] \cap [0, S]. \end{aligned} \tag{4.14}$$

Now, if $\left\|z\left(k\tilde{\Delta}\right) - z^*\left(k\tilde{\Delta}\right)\right\|_\eta$ is less than or equal to $\left(\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda\right)/\gamma(\delta, \eta, \epsilon)$, then (4.14) directly implies that

$$\|z(t) - z^*(t)\|_\eta \le \frac{\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda}{\gamma(\delta, \zeta, \nu, \epsilon)}, \qquad t \in \left[k\tilde{\Delta}, (k+1)\tilde{\Delta}\right] \cap [0, S],$$

while if $\left\|z\left(k\tilde{\Delta}\right) - z^*\left(k\tilde{\Delta}\right)\right\|_\eta$ is greater than $\left(\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda\right)/\gamma(\delta,\eta,\epsilon)$, then (4.14) and (4.10) with $t = k\tilde{\Delta}$ imply

$$\|z(t) - z^*(t)\|_\eta \le \left\|z\left(k\tilde{\Delta}\right) - z^*\left(k\tilde{\Delta}\right)\right\|_\eta \le \|z(0) - z^*(0)\|_\eta$$

for $t \in \left[k\tilde{\Delta}, (k+1)\tilde{\Delta}\right] \cap [0, S]$. We can now conclude that (4.10) holds for $t \in \left[k\tilde{\Delta}, (k+1)\tilde{\Delta}\right] \cap [0, S]$. By induction, (4.10) holds for all $t \in [0, S]$, and particularly,

$$\|z(t) - z^*(t)\|_\eta < \delta \qquad \forall t \in [0, S].$$

This implies that (4.13) holds for all $t \in [0, S]$, and by applying Corollary B.2 to the whole time domain $[0, S]$, we get the desired bound on $\|z(t) - z^*(t)\|_\eta$. ∎

Basically, Theorem 4.1 guarantees bound tracking error of C-RReg-PD provided that the algorithmic parameters are chosen to satisfy (4.5). However, we note that the right-hand side of (4.5) may be negative, meaning that the condition of Theorem 4.1 may never hold no matter how slowly the KKT point moves. To resolve this issue, we provide the following theorem.

THEOREM 4.2. *Suppose there exists $\delta > 0$ such that*

$$\Lambda_m(\delta) > M_\lambda M_{nc}(\delta). \tag{4.15}$$

Then the set of all tuples $(\beta, \eta, \epsilon) \in \mathbb{R}^3_{++}$ such that (4.5) holds is nonempty and open.

Proof. Define $f_\delta : \mathbb{R}^3_{++} \to \mathbb{R}$ by $f_\delta(\beta,\eta,\epsilon) = \delta\gamma(\delta,\eta,\epsilon) - \sqrt{\eta}\epsilon M_\lambda - \beta^{-1}\sigma_\eta$, and let

$$\eta_0 = \left(\frac{2\Lambda_m(\delta)}{\delta M_{nc}(\delta)}\right)^2, \qquad \epsilon_0 = \frac{\Lambda_m(\delta)}{\eta_0}.$$

We then have $f_\delta(\beta,\eta_0,\epsilon_0) = \frac{\delta}{2}\left(\Lambda_m(\delta) - M_\lambda M_{nc}(\delta)\right) - \beta^{-1}\sigma_{\eta_0}$. Since $\Lambda_m(\delta) > M_\lambda M_{nc}(\delta)$, we can find sufficiently large β_0 so that $f_\delta(\beta_0,\eta_0,\epsilon_0)$ is positive.

We then show that $f_\delta^{-1}(\mathbb{R}_{++})$ is open. First, we consider the map $s \mapsto \sigma_{s^{-2}}$ and show that it is convex and continuous over $s \in (0, +\infty)$. Indeed, let $t_1, t_2 \in [0, S]$ with $t_1 \ne t_2$ be given, and define

$$\overline{v}(s; t_1, t_2) \coloneqq \frac{\|z^*(t_2) - z^*(t_1)\|_{s^{-2}}}{|t_2 - t_1|} = \frac{\sqrt{\|x^*(t_2) - x^*(t_1)\|^2 + s^2\,\|\lambda^*(t_2) - \lambda^*(t_1)\|^2}}{|t_2 - t_1|}$$

for $s \in (0, +\infty)$. Obviously, $\overline{v}(s; t_1, t_2)$ is a convex function of $s \in (0, +\infty)$. Then since $\sigma_{s^{-2}}$ is the supremum of $v(s; t_1, t_2)$ over $\{(t_1, t_2) \in [0, S]^2 : t_1 \ne t_2\}$, $\sigma_{s^{-2}}$ is also a convex function of $s \in (0, +\infty)$. As $\sigma_{s^{-2}}$ is defined on an open subset of $\mathbb{R}$, we can conclude that $\sigma_{s^{-2}}$ is also a continuous function of $s \in (0, +\infty)$.

As a consequence, the function $\eta \mapsto \sigma_\eta$ is continuous, and therefore the set $f_\delta^{-1}(\mathbb{R}^3_{++})$ is an open subset of $\mathbb{R}^3_{++}$. ∎

By combining these two theorems, we get the following corollary.

COROLLARY 4.3. *Suppose (4.15) holds for some $\delta > 0$. Then there exist $\beta > 0$, $\eta > 0$, and $\epsilon > 0$ such that the bound (4.7) holds for any Lipschitz continuous solution to (3.5) whenever the initial point satisfies $\|z(0) - z^*(0)\|_\eta < \delta$. Moreover, all such parameter tuples (β, η, ϵ) form a set that has a nonempty interior in $\mathbb{R}^3_{++}$.*

We now provide some discussions on Theorems 4.1 and 4.2.

1. *Eventual tracking error bound.* The bound on the tracking error in (4.5) consists of (1) a constant term and (2) a term that decays exponentially with time t (the condition (4.5) implies $\gamma(\delta,\eta,\epsilon) < 1$). We call the constant term

$$\frac{\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda}{\gamma(\delta,\eta,\epsilon)} \tag{4.16}$$

the *eventual tracking error bound*, which can be further split into two parts:
   (a) The first part $\beta^{-1}\sigma_\eta/\gamma(\delta,\eta,\epsilon)$ is proportional to σ_η, the maximum speed of the KKT point. In time-varying optimization, such terms are common in the tracking error bound [15, 42, 36, 32].
   (b) The second part $\sqrt{\eta}\epsilon M_\lambda/\gamma(\delta,\eta,\epsilon)$ is proportional to M_λ, the maximum magnitude of the optimal dual variable $\lambda^*(t)$. This term represents the additional error introduced by dual regularization. We mention that this term cannot be eliminated by setting $\epsilon = 0$, as this would make $\gamma(\delta,\eta,\epsilon)$ negative, violating the condition (4.5).[2]
2. *Interpretation of Theorem* 4.2 *and the condition* (4.15). Theorem 4.2 provides a sufficient condition for guaranteeing the existence of algorithmic parameters such that the tracking error bound in Theorem 3.1 will apply; such algorithmic parameters (β,η,ϵ) will be called *feasible algorithmic parameters*. The openness of the set of feasible parameters then indicates that the choice of algorithmic parameters is robust to small perturbations.
   Loosely speaking, the condition (4.15) can be interpreted as requiring that the problem (1.1) should be sufficiently convex around the optimal trajectory to overcome the nonlinearity of the nonconvex constraints. We further give two examples for which (4.15) holds:
   (a) When $f^{nc}(x,t) = 0$ and $c(\cdot,t)$ is μ-strongly convex for all t for some $\mu > 0$, we have $\Lambda_m(\delta) \geq \mu$ and $M_{nc}(\delta) = 0$ for all $\delta > 0$. Therefore, (4.15) automatically holds for any $\delta > 0$, and (4.5) can be satisfied for any parameter tuple (β,η,ϵ) by choosing δ to be sufficiently large. Theorem 4.1 then implies bounded tracking error of C-RReg-PD. This can be regarded as the continuous-time analogue of the results in [38]. However, we should emphasize that [38] only shows bounded tracking error with respect to an *approximate* KKT point, while our result establishes bounded tracking error with respect to a true KKT trajectory.
   (b) We still suppose $f^{nc}(x,t) = 0$, i.e., the constraints are convex, but for the cost function c, we only assume that $\nabla^2_{xx}c(x^*(t),t)$ is positive definite for each t. In this case, we have $\Lambda_m(0) > 0$ and $M_{nc}(\delta) = 0$ for all $\delta > 0$. Then, by Assumption 2.1 and Berge's maximum theorem, it can be shown that $\Lambda_m(\delta)$ is a continuous function defined over $\delta \geq 0$. Thus there exists $\delta > 0$ such that $\Lambda_m(\delta) > 0$. Consequently, by Corollary 4.3, there exist $\beta > 0$, $\eta > 0$, and $\epsilon > 0$ such that C-RReg-PD enjoys bounded tracking error whenever $\|z(0) - z^*(0)\|_\eta < \delta$.

4.1. Isolation of the KKT trajectory. In section 2, we remarked that there could be multiple trajectories of KKT points, and in Theorem 4.1, the tracking performance of C-RReg-PD was analyzed with respect to one of these trajectories that is arbitrarily chosen. On the other hand, if the KKT trajectory $z^*(t)$ bifurcates into

[2]However, this does not mean that RReg-PD or C-RReg-PD would not work by setting $\epsilon = 0$. We leave theoretical explanation for the $\epsilon = 0$ case as future work.

two or more branches at some time $\tilde{t} \in [0, S]$ and these branches become far away as time proceeds, then it appears that we are not able to identify from Theorem 4.1 which trajectory the algorithm will track. Fortunately, as the following theorems show, under certain conditions, such possibilities will not occur.

THEOREM 4.4. *Suppose for some $\delta > 0$ and $\eta > 0$,*

$$\Lambda_m(\delta) - \frac{\sqrt{\eta}}{2}\delta M_{nc}(\delta) > 0. \tag{4.17}$$

Then there is no KKT point in the set $\{z = (x, \lambda) : 0 < \|z - z^*(t)\|_\eta \le \delta,\ x \ne x^*(t)\}$ for each $t \in [0, S]$.

In particular, (4.17) *holds if the condition* (4.5) *holds for some $\delta \le 2\eta^{-1/2} M_\lambda$.*

The proof of Theorem 4.4 is given in Appendix C. Some discussions are as follows.

1. Theorem 4.4 shows that if the condition of Theorem 4.1 holds for some $\delta \le 2\eta^{-1/2} M_\lambda$, then within the neighborhood $0 < \|z - z^*(t)\|_\eta \le \delta$, there does not exist a KKT point whose primal variable is different from $x^*(t)$. This implies that merging or bifurcation of KKT trajectories does not occur when the condition of Theorem 4.1 holds within a sufficiently small neighborhood around the KKT point.
2. We mention that Theorem 4.4 does not exclude the possibility that at time t, there exists $\lambda^+ \ne \lambda^*(t)$ such that $(x^*(t), \lambda^+)$ is also a KKT point of (1.1) and $\|(x^*(t), \lambda^+) - (x^*(t), \lambda^*)\|_\eta \le \delta$, unless we also assume that the optimal dual variable associated with $x^*(t)$ is unique at time t. A typical constraint qualification that guarantees the uniqueness of the optimal Lagrange multiplier is the linear independence constraint qualification (LICQ) [45], which cannot be directly used in our setting as the set $\mathcal{X}(t)$ is not explicitly specified by inequalities, but can be possibly checked if we can express $x \in \mathcal{X}(t)$ by a group of inequality constraints.

5. Choice of parameters. In this section, we provide some brief discussions on the choice of the algorithmic parameters.

First, we note that the eventual tracking error bound decreases monotonically as β increases, which appears to suggest that β should be as large as possible. On the other hand, we point out that, in practice, since the original RReg-PD (1.2) is implemented in discrete time, a large β will then lead to a large step size $\alpha_T = \beta\Delta_T$ that can lead to oscillations or instability. A rigorous discussion on the choice of β can be complicated and is out of the scope of this paper.

We now assume that $\beta > 0$ is given and fixed. The following theorem suggests a way to choose the regularization parameter ϵ based on η, whose proof is in Appendix D.

THEOREM 5.1. *Denote*

$$\mathscr{A}_{\text{fp}}(\delta, \beta) := \left\{(\eta, \epsilon) > 0 : \beta^{-1}\sigma_\eta < \delta\,\gamma(\delta, \eta, \epsilon) - \sqrt{\eta}\epsilon M_\lambda\right\}.$$

Let $\delta > 0$ and $\beta > 0$ be fixed such that $\mathscr{A}_{\text{fp}}(\delta, \beta) \ne \varnothing$. Then, we have the following:

1. *For any fixed $\eta > 0$, the choice $\epsilon = \Lambda_m(\delta)/\eta$ minimizes the eventual tracking error bound* (4.16) *over $\{\epsilon : (\eta, \epsilon) \in \mathscr{A}_{\text{fp}}(\delta, \beta)\}$.*
2. *There exists $\eta^* > 0$ such that $(\eta^*, \Lambda_m(\delta)/\eta^*)$ minimizes the eventual tracking error bound* (4.16) *over $\mathscr{A}_{\text{fp}}(\delta, \beta)$.*

Theorem 5.1 suggests that we can choose ϵ such that the product $\eta\epsilon$ approximates the local convexity of the problem (1.1). In practice, if we have some prior knowledge

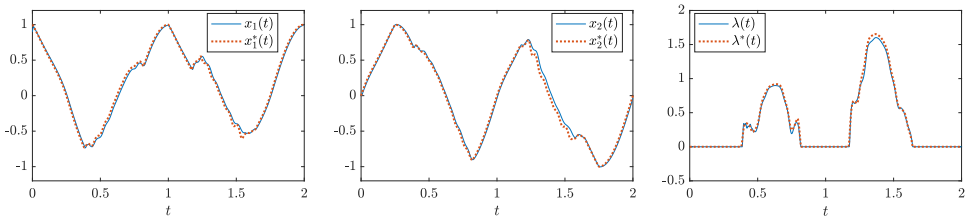


FIG. 1. *The trajectories* $z(t) = (x_1(t), x_2(t), \lambda(t))$ *and* $z^*(t) = (x_1^*(t), x_2^*(t), \lambda^*(t))$.

of the functions $c(x,t)$ and $f(x,t)$, so that some rough estimate of $\Lambda_m(\delta)$ is available, one can then use this prior information for choosing ϵ based on η. After fixing the relation between ϵ and η, the choice of η can then be determined by carrying out a parameter sweep.

On the other hand, we should also point out that the choice $\eta\epsilon = \Lambda_m(\delta)$ is based on analyzing the eventual tracking error bound (4.16), which is a conservative estimate of the real tracking error. Nevertheless, the analysis presented here will still be of value and can serve as a guide for choosing the parameters in practice. In section 6, we will provide some numerical justification for this choice.

6. Numerical examples. In this section, we test the performance of C-RReg-PD on a numerical test case. As Theorem 3.1 shows, we can simulate the C-RReg-PD dynamics by the discrete-time iteration (1.2) with $\alpha = \beta\Delta_T$ for a sufficiently small Δ_T.

The time-varying optimization problem of the test case is formulated as

$$\begin{aligned} \min_{x\in\mathcal{X}(t)} \quad & \frac{1}{2}(x-r(t))^T \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix} (x-r(t)) + \frac{1-\cos(8\|x-r(t)\|)}{32} \\ \text{s.t.} \quad & \frac{\cos^2(50t)}{5} x_1 x_2 - \sin(\pi t)\, x_1 - x_2 \le 0, \end{aligned}$$

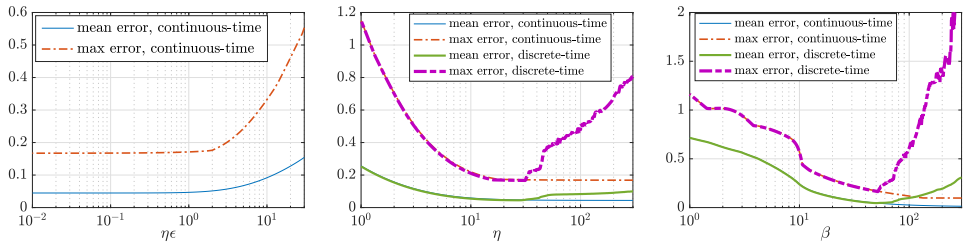
where

$$\mathcal{X}(t) = \left\{ x \in \mathbb{R}^2 : \|x\| \le 1 - \frac{1}{4}\sin(4\pi t) \right\}, \qquad r(t) = \begin{bmatrix} \cos(2\pi t) \\ \sin(2\pi t) \end{bmatrix}$$

for each $t \in [0,2]$. It can be checked that the objective function and the inequality constraint are nonconvex over $x \in \mathbb{R}^2$ for each $t \in [0,2]$, but we have $\Lambda_m(\delta) \ge 1$ and $M_{nc}(\delta) = 1/5$ for any $\delta \ge 0$. We simulate both the discrete-time RReg-PD and C-RReg-PD. We fix $\beta = 50$ for C-RReg-PD, and let $\alpha = \beta\Delta_T$ for the discrete-time RReg-PD. The initial point is $(x_1(0), x_2(0), \lambda(0)) = (0.9, 0, 0)$.

Figure 1 shows the trajectory of C-RReg-PD $z(t) = (x_1(t), x_2(t), \lambda(t))$ and a locally optimal trajectory $z^*(t)$, where we set $\eta = 30$ and $\epsilon = 1/\eta$. We can see that C-RReg-PD is able to track the optimal trajectory with reasonable bounded tracking error; the average and maximum tracking error are given by 0.0468 and 0.1712, respectively.

Next, we fix $\eta = 30$ and vary ϵ, and the results are shown in Figure 2a. It can be seen that for $\epsilon \gtrsim \Lambda_m(\delta)/\eta$, the tracking error increases as ϵ increases. However, for $\epsilon \lesssim \Lambda_m(\delta)/\eta$, the tracking error still decreases as ϵ decreases, which is contrary to the analysis in section 5. We suspect that this is due to some artifact in the analysis of the eventual tracking error bound for very small ϵ. Nevertheless, the choice $\epsilon = 1/\eta \approx \Lambda_m(\delta)/\eta$ gives almost indistinguishable tracking error compared to smaller values of ϵ, which suggests that it is still a good choice in practice.



(a) Tracking error for different values of $\eta\epsilon$.

(b) Tracking error for different values of η.

(c) Tracking error for different values of β.

FIG. 2. *Tracking error for different choices of parameters.*

We then fix $\epsilon = 1/\eta$ and run parameter sweeps for η and β, where we fix $\beta = 50$ for the sweep of η and fix $\eta = 30$ for the sweep of β. We also add simulation of the discrete-time RReg-PD with sampling interval $\Delta_T = 1 \times 10^{-3}$ and step size $\alpha = \beta\Delta_T$. The results are shown in Figures 2b and 2c. Some observations and discussions are as follows:

1. For C-RReg-PD, the tracking performance improves as we increase β, which is expected from the (eventual) tracking error bound (4.16). On the other hand, the tracking error decreases monotonically as we increase η; this is contrary to the second part of Theorem 5.1, and we again suspect that this is due to some artifact in our analysis for small ϵ. Improvement of the tracking error bound will be an interesting future direction.
2. For the discrete-time RReg-PD, when η and β are below certain thresholds, the tracking performance is very close to C-RReg-PD. This suggests that the study of C-RReg-PD can indeed provide insights on the performance of the discrete-time RReg-PD algorithm. But if η or β is chosen to be too large, then there can be a significant gap of the tracking errors between the discrete-time and the continuous-time algorithms. Recalling that $\alpha = \beta\Delta_T$ and $\eta\alpha = \eta\beta\Delta_T$ are the step sizes of the primal and the dual updates, the observed phenomenon is expected as Theorem 3.1 indicates that we can use C-RReg-PD to approximate the discrete-time RReg-PD only when the step sizes are small.

7. Conclusion and future direction. In this paper, we studied the regularized primal-dual gradient method for time-varying nonconvex optimization by analyzing its continuous-time counterpart. We derived sufficient conditions that guarantee bounded tracking error for the continuous-time algorithm. We also studied conditions under which the KKT trajectories will be isolated, and investigated the optimal choice of algorithmic parameters. Implications of these analytical results were discussed, and a numerical example was presented to illustrate the performance of the proposed algorithm. Some possible generalizations and directions to explore include: (1) studying how to handle jumps in the KKT trajectory $z^*(t)$, (2) investigating other metrics for evaluating the tracking error, to see whether weaker conditions for bounded tracking error can be derived, (3) adapting techniques for escaping saddle-points for gradient methods in the time-varying setting, and (4) investigating classes of problems for which the conditions of bounded tracking error can be verified and the algorithmic parameters can be tuned more easily.

Appendix A. Proof of Theorem 3.1. Theorem 3.1 follows from the following lemma on perturbed sweeping process and the catching algorithm [10, 11].

LEMMA A.1. *Suppose* $\Psi : \mathbb{R}^p \times [0,S] \to \mathbb{R}^p$ *and* $\mathcal{C} : [0,S] \to 2^{\mathbb{R}^p}$, *and that*

1. $\mathcal{C}$ *is* κ_1*-Lipschitz, and for each* $t \in [0,S]$, $\mathcal{C}(t)$ *is closed and convex,*
2. Ψ *is continuous when restricted to the set* $\bigcup_{t\in[0,S]} \mathcal{C}(t)\times[0,S]$, *and there exists some* $\kappa_2 > 0$ *such that* $\|\Psi(z,t)\| \le \kappa_2(1+\|z\|)$ *for all* $(z,t) \in \bigcup_{t\in[0,S]} \mathcal{C}(t) \times [0,S]$.

Let $z_0 \in \mathcal{C}(0)$ *be arbitrary, and for each* $T \in \mathbb{N}$, *define* $z_\tau^{(T)}$, $\tau \in \{0,1,2,\dots,T\}$, *by*

$$z_0^{(T)} = z_0, \qquad z_\tau^{(T)} = \mathcal{P}_{\mathcal{C}(\tau\Delta_K)}\left[z_{\tau-1}^{(T)} + \Delta_T \Psi\left(z_{\tau-1}^{(T)}, \tau\Delta_T\right)\right],$$

where $\Delta_T := S/T$, *and for* $t \in [0,S]$, *define*

$$z^{[T]}(t) = \frac{\tau\Delta_T - t}{\Delta_T} z_{\tau-1}^{(T)} + \frac{t-(\tau-1)\Delta_T}{\Delta_T} z_\tau^{(T)} \tag{A.1}$$

if $t \in [(\tau-1)\Delta_T, \tau\Delta_T]$. *Then, if we keep* S *constant and let* $T \to \infty$, *the sequence* $\left(z^{[T]}\right)_{T\in\mathbb{N}}$ *defined in* (A.1) *has a convergent subsequence, and any convergent subsequence converges uniformly to a Lipschitz continuous* $z(t), t \in [0,S]$, *that satisfies*

$$z(0) = z_0 \quad \textit{and} \quad -\frac{d}{dt}z(t) + \Psi(z(t),t) \in N_{\mathcal{C}(t)}(z(t)) \quad \forall t \in [0,S] \text{ a.e.} \tag{A.2}$$

It is straightforward to see that Theorem 3.1 follows from Lemma A.1 by letting $\Psi(z,t) = \beta\Phi(z,t)$ and checking each of the conditions of Lemma A.1.

1. $\mathcal{C}$ is κ_1-Lipschtz as $\mathcal{X}$ is κ_1-Lipschtz.
2. Ψ is obviously continuous on $\bigcup_{t\in[0,S]} \mathcal{X}(t) \times [0,S]$. Moreover,

$$\|\Psi(z,t)\| \le \beta\left(\|\nabla_x c(x,t)\| + \|J_{f,x}(x,t)\|\,\|\lambda\|\right) + \eta\beta\,\|f(x,t)\| + \eta\beta\epsilon\,\|\lambda\|.$$

Let $x_{\text{aux}} \in \bigcup_{t\in[0,S]} \mathcal{X}(t)$ be arbitrary, and

$$\kappa_3 := \sup\left\{\|J_{f,x}(x,t)\| : (x,t) \in \bigcup_{t\in[0,S]} \mathcal{X}(t)\times[0,S]\right\},$$

$$\kappa_4 := \sup_{t\in[0,S]} \|f(x_{\text{aux}},t)\|,$$

both of which are finite. Then $\|f(x,t)\| \le \kappa_4 + \kappa_3(\|x\| + \|x_{\text{aux}}\|)$, and by (3.4) and noticing that $\|x\| \le \|z\|$ and $\|\lambda\| \le \|z\|$, we get

$$\begin{aligned}\|\Psi(z,t)\| &\le \beta(\kappa_2(1+\|z\|) + \kappa_3\|z\|) + \eta\beta\left(\kappa_4 + \kappa_3(\|z\| + \|x_{\text{aux}}\|)\right) + \eta\beta\epsilon\|z\| \\ &\le \kappa_5(1+\|z\|)\end{aligned}$$

for some $\kappa_5 > 0$.

By Lemma A.1, the sequence of trajectories defined by (A.1)—and consequently (3.2)—then has convergent subsequences each of which converges to some Lipschitz continuous solution to (A.2).

Appendix B. Auxiliary results for the proof of Theorem 4.1. We first provide the following Gronwall-type lemma.

LEMMA B.1 (see [46]). *Let I be a closed interval with zero as left endpoint. Let $u(t)$ be a continuous nonnegative function that satisfies $u(t) \le u_0 + \int_0^t w(s)u^p(s)\,ds$, where $w(t)$ is a continuous nonnegative function on I. Then for $0 \le p < 1$, we have*

$$u(t) \le \left(u_0^{1-p} + (1-p)\int_0^t w(s)\,ds\right)^{1/(1-\beta)}.$$

Based on Lemma B.1, we derive a corollary that can be directly used in our proof.

COROLLARY B.2. *Let $v(t)$ be a nonnegative absolutely continuous function satisfying $\frac{1}{2}\frac{d}{dt}\big(v^2(t)\big) \le av(t) - bv^2(t)$ for almost all $t \in [0,S]$, where $a \ge 0$ and $b > 0$ are constants. Then*

$$v(t) \le e^{-bt}v(0) + \frac{a}{b}(1 - e^{-bt}).$$

Proof of Corollary B.2. Define $u(t) = e^{2\beta t}v^2(t)$. Then it can be checked that

$$\dot{u}(t) = 2\beta e^{2\beta t}v^2(t) + e^{2\beta t}\frac{d\big(v^2(t)\big)}{dt} \le 2\beta e^{2\beta t}v^2(t) + 2e^{2\beta t}(\alpha v(t) - \beta v^2(t)) = 2\alpha e^{\beta t}\sqrt{u(t)}$$

for almost all $t \in [0,S]$. Therefore, by Lemma B.1,

$$u(t) \le \left(\sqrt{u(0)} + \alpha\int_0^t e^{\beta s}\,ds\right)^2 = \left(\sqrt{u(0)} + \frac{\alpha}{\beta}(e^{\beta t} - 1)\right)^2,$$

and by the definition of $u(t)$, we get the desired result. ∎

Appendix C. Proof of Theorem 4.4. We shall frequently denote $z(t) = z = (x,\lambda)$ and $z^*(t) = z^* = (x^*,\lambda^*)$, where the dependence on t is suppressed for simplicity. Suppose for some $t \in [0,S]$, there is another KKT point $z^+ = (x^+,\lambda^+)$ satisfying $0 < \|z^+ - z^*\|_\eta \le \delta$ and $x^+ \ne x^*$. By (2.2b), we have $0 \ge (x^+ - x^*)^T\big(-\nabla_x \mathcal{L}^{nc}(x^*,\lambda^*,t) - J_{f^c,x}(x^*,t)^T\lambda^*\big)$ and

$$0 \ge (x^* - x^+)^T\big(-\nabla_x\mathcal{L}^{nc}(x^+,\lambda^*,t) - J_{f^c,x}(x^+,t)^T\lambda^+ + J_{f^{nc},x}(x^+,t)^T\big(\lambda^* - \lambda^+\big)\big).$$

Thus,

$$\begin{aligned}
0 \ge{}& (x^+ - x^*)^T\big(\nabla_x\mathcal{L}^{nc}(x^+,\lambda^*,t) - \nabla_x\mathcal{L}^{nc}(x^*,\lambda^*,t)\\
&+ J_{f^c,x}(x^+,t)^T\lambda^+ - J_{f^c,x}(x^*,t)^T\lambda^*(t) + J_{f^{nc},x}(x^+,t)^T\big(\lambda^+ - \lambda^*\big)\big)\\
={}& (x^+ - x^*)^T B_{\mathcal{L}^{nc}}(x^+ - x^*) + (x^+ - x^*)^T\left(J_{f^c,x}(x^+,t)^T\lambda^+ - J_{f^c,x}(x^*,t)\lambda^*\right)\\
&+ (x^+ - x^*)^T J_{f^{nc},x}(x^+,t)^T\big(\lambda^+ - \lambda^*\big),
\end{aligned}$$

where $B_{\mathcal{L}^{nc}} := \overline{H}_{\mathcal{L}^{nc}}(x^+ - x^*,t)$. Also, by the complementary slackness condition, $(\lambda^+ - \lambda^*)^T\left(f(x^+,t) - f(x^*,t)\right) = -\lambda^{+T}f(x^*,t) - \lambda^{*T}f(x^+,t) \ge 0$. Therefore,

$$\begin{aligned}
0 \ge{}& (x^+ - x^*)^T B_{\mathcal{L}^{nc}}(x^+ - x^*) + (x^+ - x^*)^T\left(J_{f^c,x}(x^+,t)^T\lambda^+ - J_{f^c,x}(x^*,t)\lambda^*(t)\right)\\
&+ \left(\lambda^+ - \lambda^*\right)^T\left(J_{f^{nc},x}(x^+,t)(x^+ - x^*) - \big(f(x^+,t) - f(x^*,t)\big)\right).
\end{aligned}$$

Notice that

$$(x^+ - x^*)^T\left(J_{f^c,x}(x^+,t)^T\lambda^+ - J_{f^c,x}(x^*,t)\lambda^*\right)$$

$$
\begin{aligned}
&+ \left(\lambda^+ - \lambda^*\right)^T \left(J_{f^{nc},x}(x^+,t)(x^+ - x^*) - \left(f(x^+,t) - f(x^*,t)\right)\right) \\
&= \lambda^{+T} \left(f^c(x^*,t) - f^c(x^+,t) - J_{f^c,x}(x^+,t)(x^* - x^+)\right) \\
&\quad + \lambda^{*T} \left(f^c(x^+,t) - f^c(x^*,t) - J_{f^c,x}(x^*,t)(x^+ - x^*)\right) \\
&\quad - \left(\lambda^+ - \lambda^*\right)^T \left(f^{nc}(x^+,t) + J_{f^{nc},x}(x^+,t)(x^* - x^+) - f^{nc}(x^*,t)\right) \\
&\geq (x^+ - x^*)^T \left(\frac{1}{2}\sum\nolimits_{i=1}^m \lambda_i^* B_{f_i^c}\right)(x^+ - x^*) \\
&\quad - \left(\lambda^+ - \lambda^*\right)^T \left(f^{nc}(x^+,t) + J_{f^{nc},x}(x^+,t)(x^* - x^+) - f^{nc}(x^*,t)\right),
\end{aligned}
$$

where we denote $B_{f_i^c} := \overline{H}_{f_i^c}(x^+ - x^*, t)$. Therefore,

$$
\begin{aligned}
0 &\geq (x^+ - x^*)^T \left(B_{\mathcal{L}^{nc}} + \frac{1}{2}\sum\nolimits_{i=1}^m \lambda_i^* B_{f_i^c}\right)(x^+ - x^*) \\
&\quad - \left(\lambda^+ - \lambda^*\right)^T \left(f^{nc}(x^+,t) + J_{f^{nc},x}(x^+,t)(x^* - x^+) - f^{nc}(x^*,t)\right) \\
&\geq \Lambda_m(\delta)\|x^+ - x^*\|^2 - \frac{M_{nc}(\delta)}{2}\|x^+ - x^*\|^2\left\|\lambda^+ - \lambda^*\right\| \\
&\geq \left(\Lambda_m(\delta) - \frac{\sqrt{\eta}}{2}\delta M_{f^{nc}}(\delta)\right)\|x^+ - x^*\|^2.
\end{aligned}
$$

However, if (4.17) holds, the right-hand side of the above inequality is then positive, leading to a contradiction.

Now we prove that (4.17) holds if (4.5) holds for some $\delta \leq 2\eta^{-1/2}M_\lambda$. We have $\min\{\Lambda_m(\delta), \eta\epsilon\} \leq \frac{1}{2}(\Lambda_m(\delta) + \eta\epsilon)$, and so $\Lambda_m(\delta) \geq 2\min\{\Lambda_m(\delta), \eta\epsilon\} - \eta\epsilon$. On the other hand, (4.5) implies that

$$
\min\{\Lambda_m(\delta), \eta\epsilon\} > \frac{\sqrt{\eta}}{4}\delta M_{nc}(\delta) + \delta^{-1}\left(\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda\right) > \frac{\sqrt{\eta}}{4}\delta M_{nc}(\delta) + \delta^{-1}\sqrt{\eta}\epsilon M_\lambda.
$$

Since $\delta \leq 2\eta^{-1/2}M_\lambda$, we have $\delta^{-1}\sqrt{\eta}\epsilon M_\lambda \geq \eta\epsilon/2$, and so

$$
\Lambda_m(\delta) > 2\left(\frac{\sqrt{\eta}}{4}\delta M_{nc}(\delta) + \delta^{-1}\sqrt{\eta}\epsilon M_\lambda\right) - \eta\epsilon \geq \frac{\sqrt{\eta}}{2}\delta M_{nc}(\delta).
$$

This completes the proof. ∎

Appendix D. Proof of Theorem 5.1. Given δ, β, denote

$$
g_0(\eta,\epsilon) = \frac{\beta^{-1}\sigma_\eta + \sqrt{\eta}\epsilon M_\lambda}{\gamma(\delta,\eta,\epsilon)}, \qquad g_1(\eta,\epsilon) = \delta\,\gamma(\delta,\eta,\epsilon) - \sqrt{\eta}\epsilon M_\lambda - \beta^{-1}\sigma_\eta
$$

for $\eta > 0, \epsilon > 0$. Obviously, $g_0(\eta,\epsilon)$ is just the eventual tracking error bound, and $g_0(\eta,\epsilon) < \delta$ if $g_1(\eta,\epsilon) > 0$.

First, let $\eta > 0$ be fixed. Then $g_1(\eta,\epsilon)$ is a continuous function of ϵ, is monotonic when $\epsilon \leq \Lambda_m(\delta)/\eta$, and is decreasing when $\epsilon \geq \Lambda_m(\delta)/\eta$. Thus, the set $\mathscr{I}_\epsilon = \{\epsilon > 0 : (\eta,\epsilon) \in \mathscr{A}_{\text{fp}}(\delta,\beta)\} = \{\epsilon > 0 : g_1(\eta,\epsilon) > 0\}$ is an open interval. We then have

$$
\frac{\partial}{\partial\epsilon}g_0(\eta,\epsilon) = \begin{cases} -\dfrac{\eta\left(\delta M_\lambda M_{nc}(\delta)/4 + \beta^{-1}\sigma_\eta\right)}{\left(\eta\epsilon - \sqrt{\eta}\delta M_{nc}(\delta)/4\right)^2} \leq 0, & \epsilon \in \mathscr{I}_\epsilon \cap (0, \Lambda_m(\delta)/\eta), \\ \dfrac{\sqrt{\eta}M_\lambda}{\Lambda_m(\delta) - \sqrt{\eta}\delta M_{nc}(\delta)/4} \geq 0, & \epsilon \in \mathscr{I}_\epsilon \cap (\Lambda_m(\delta)/\eta, +\infty). \end{cases}
$$

Therefore, $\epsilon = \Lambda_m(\delta)/\eta$ minimizes the eventual tracking error bound over $\{\epsilon > 0 : (\eta,\epsilon) \in \mathscr{A}_{\text{fp}}(\delta,\beta)\}$.

We now define the functions

$$\tilde{g}_0(x) = g_0\big(x^{-2}, x^2\Lambda_m(\delta)\big) = \frac{\beta^{-1}\sigma_{x^{-2}} + x\Lambda_m(\delta)M_\lambda}{\Lambda_m(\delta) - x^{-1}\delta M_{nc}(\delta)/4}, \quad \tilde{g}_1(x) = g_1\big(x^{-2}, x^2\Lambda_m(\delta)\big)$$

for $x \in (x_l, +\infty)$, where $x_l := \frac{\delta M_{nc}(\delta)}{4\Lambda_m(\delta)}$. Then a minimizer x^* of $\tilde{g}_0(x)$ leads to a minimizer $(\eta^*, \epsilon^*) = (x^{*-2}, x^{*2}\Lambda_m(\delta))$ of $g_0(\eta, \epsilon)$, and we need only prove that $\tilde{g}_0(x)$ has a minimizer over $(x_l, +\infty)$.

Since $\sigma_{x^{-2}}$ is a convex and nondecreasing function of x (see the proof of Theorem 4.2), it is absolutely continuous and admits a weak derivative $u(x)$ which is nonnegative and nondecreasing in x. Then we see that $\tilde{g}_0(x)$ is also absolutely continuous, whose weak derivative is given by

$$D\tilde{g}_0(x) = \frac{\beta^{-1}u(x) + \Lambda_m(\delta)M_\lambda - \frac{\beta^{-1}x^{-1}\sigma_{x^{-2}} + \Lambda_m(\delta)M_\lambda}{x(\Lambda_m(\delta) - x^{-1}\delta M_{nc}(\delta)/4)}\frac{\delta M_{nc}(\delta)}{4}}{\Lambda_m(\delta) - x^{-1}\delta M_{nc}(\delta)/4}.$$

Since $x^{-1}\sigma_{x^{-2}}$ is equal to

$$\sup\left\{\frac{(x^{-2}\left\|x^*(t_2) - x^*(t_1)\right\|^2 + \left\|\lambda^*(t_2) - \lambda^*(t_1)\right\|^2)^{1/2}}{|t_2 - t_1|} : t_1, t_2 \in [0, S], t_1 \neq t_2\right\},$$

we see that $x^{-1}\sigma_{x^{-2}}$ is nonincreasing in x. Then we can easily verify that

$$\frac{\beta^{-1}x^{-1}\sigma_{x^{-2}} + \Lambda_m(\delta)M_\lambda}{x(\Lambda_m(\delta) - x^{-1}\delta M_{nc}(\delta)/4)}$$

is a strictly decreasing function of x, as the numerator is nonincreasing in x and the denominator is positive and strictly increasing in x for $x > \delta M_{nc}(\delta)/(4\Lambda_m(\delta))$. Moreover, we have

$$\begin{aligned}
&\lim_{x\to+\infty}\Big(\beta^{-1}u(x) + \Lambda_m(\delta)M_\lambda - \frac{\beta^{-1}x^{-1}\sigma_{x^{-2}} + \Lambda_m(\delta)M_\lambda}{x(\Lambda_m(\delta) - x^{-1}\delta M_{nc}(\delta)/4)}\frac{\delta M_{nc}(\delta)}{4}\Big)\\
&\quad = \beta^{-1}\lim_{x\to+\infty}u(x) + \Lambda_m(\delta)M_\lambda > 0,\\
&\lim_{x\to x_l^+}\Big(\beta^{-1}u(x) + \Lambda_m(\delta)M_\lambda - \frac{\beta^{-1}x^{-1}\sigma_{x^{-2}} + \Lambda_m(\delta)M_\lambda}{x(\Lambda_m(\delta) - x^{-1}\delta M_{nc}(\delta)/4)}\frac{\delta M_{nc}(\delta)}{4}\Big) = -\infty.
\end{aligned}$$

Therefore, there exists $x^* \in (x_l, +\infty)$ such that $D\tilde{g}_0(x) \le 0$ for $x \in (x_l, x^*)$ and $D\tilde{g}_0(x) \ge 0$ for $x \in (x^*, +\infty)$. We then see that $\tilde{g}_0(x)$ has a minimizer x^* over $(x_l, +\infty)$, which completes the proof. ∎

REFERENCES

[1] S. Adly, T. Haddad, and L. Thibault, *Convex sweeping process in the framework of measure differential inclusions and evolution variational inequalities*, Math. Program., 148 (2014), pp. 5–47.

[2] Z. Allen-Zhu, *Natasha 2: Faster non-convex optimization than SGD*, in Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Curran Associates, Red Hook, NY, 2018.

[3] A. Balavoine, C. J. Rozell, and J. Romberg, *Discrete and continuous-time soft-thresholding for dynamic signal recovery*, IEEE Trans. Signal Process., 63 (2015), pp. 3165–3176.

[4] A. Bernstein and E. Dall'Anese, *Asynchronous and distributed tracking of time-varying fixed points*, in Proceedings of the 57th IEEE Conference on Decision and Control, IEEE, 2018, pp. 3236–3243.

[5] A. Bernstein, E. Dall'Anese, and A. Simonetto, *Online primal-dual methods with measurement feedback for time-varying convex optimization*, IEEE Trans. Signal Process., 67 (2019), pp. 1978–1991.
[6] G. Bianchin and F. Pasqualetti, *Gramian-based optimization for the analysis and control of traffic networks*, IEEE Trans. Intell. Transp. Syst., 21 (2020), pp. 3013–3024.
[7] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[8] B. Brogliato, A. Daniilidis, C. Lemaréchal, and V. Acary, *On the equivalence between complementarity systems, projected systems and differential inclusions*, Systems Control Lett., 55 (2006), pp. 45–51.
[9] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, *Accelerated methods for nonconvex optimization*, SIAM J. Optim., 28 (2018), pp. 1751–1772, https://doi.org/10.1137/17M1114296.
[10] C. Castaing, T. X. Dúc Hā, and M. Valadier, *Evolution equations governed by the sweeping process*, Set-Valued Anal., 1 (1993), pp. 109–139.
[11] C. Castaing and M. D. P. Monteiro Marques, *BV periodic solutions of an evolution problem associated with continuous moving convex sets*, Set-Valued Anal., 3 (1995), pp. 381–399.
[12] J. Chen and V. K. N. Lau, *Convergence analysis of saddle point problems in time varying wireless systems—control theoretical approach*, IEEE Trans. Signal Process., 60 (2012), pp. 443–452.
[13] A. Cherukuri, E. Mallada, S. Low, and J. Cortés, *The role of convexity on saddle-point dynamics: Lyapunov function and robustness*, IEEE Trans. Automat. Control, 63 (2018), pp. 2449–2464.
[14] P. Cisneros-Velarde, S. Jafarpour, and F. Bullo, *A contraction analysis of primal-dual dynamics in distributed and time-varying implementations*, IEEE Trans. Automat. Control, 67 (2022), pp. 3560–3566, https://doi.org/10.1109/TAC.2021.3103865.
[15] E. Dall'Anese and A. Simonetto, *Optimal power flow pursuit*, IEEE Trans. Smart Grid, 9 (2018), pp. 942–952.
[16] Y. Ding, J. Lavaei, and M. Arcak, *Escaping spurious local minimum trajectories in online time-varying nonconvex optimization*, in Proceedings of the 2021 American Control Conference, IEEE, 2021, pp. 454–461.
[17] R. Dixit, A. S. Bedi, R. Tripathi, and K. Rajawat, *Online learning with inexact proximal online gradient descent algorithms*, IEEE Trans. Signal Process., 67 (2019), pp. 1338–1352.
[18] A. L. Dontchev, M. I. Krastanov, R. T. Rockafellar, and V. M. Veliov, *An Euler–Newton continuation method for tracking solution trajectories of parametric variational inequalities*, SIAM J. Control Optim., 51 (2013), pp. 1823–1840, https://doi.org/10.1137/120876915.
[19] A. L. Dontchev and R. T. Rockafellar, *Implicit Functions and Solution Mappings*, 2nd ed., Springer, New York, 2014.
[20] H. Feng, H. Zhang, and J. Lavaei, *A dynamical system perspective for escaping sharp local minima in equality constrained optimization problems*, in Proceedings of the 59th IEEE Conference on Decision and Control, IEEE, 2020, pp. 4255–4261.
[21] X. Gao, X. Li, and S. Zhang, *Online learning with non-convex losses and non-stationary regret*, in Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 235–243.
[22] B. Gharesifard and J. Cortés, *Distributed continuous-time convex optimization on weight-balanced digraphs*, IEEE Trans. Automat. Control, 59 (2014), pp. 781–786.
[23] J. Guddat, F. Guerra Vazquez, and H. Th. Jongen, *Parametric Optimization: Singularities, Pathfollowing and Jumps*, B.G. Teubner, Stuttgart, 1990.
[24] A. Hauswirth, I. Subotić, S. Bolognani, G. Hug, and F. Dörfler, *Time-varying projected dynamical systems with applications to feedback optimization of power systems*, in Proceedings of the 57th IEEE Conference on Decision and Control, IEEE, 2018, pp. 3258–3263.
[25] A. Jadbabaie, A. Rakhlin, S. Shahrampour, and K. Sridharan, *Online optimization: Competing with dynamic comparators*, in Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, PMLR, 2015, pp. 398–406.
[26] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, *How to escape saddle points efficiently*, in Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017, pp. 1724–1732.
[27] J. Koshal, A. Nedić, and U. V. Shanbhag, *Multiuser optimization: Distributed algorithms and error analysis*, SIAM J. Optim., 21 (2011), pp. 1046–1081, https://doi.org/10.1137/090770102.
[28] A. Lesage-Landry, J. A. Taylor, and I. Shames, *Second-order online nonconvex optimiza-*

tion, IEEE Trans. Automat. Control, 66 (2021), pp. 4866–4872.
[29] Q. Ling, Y. Xu, W. Yin, and Z. Wen, *Decentralized low-rank matrix completion*, in Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2012, pp. 2925–2928.
[30] S. H. Low and D. E. Lapsley, *Optimization flow control—*I*: Basic algorithm and convergence*, IEEE/ACM Trans. Netw., 7 (1999), pp. 861–874.
[31] J. R. Marden, S. D. Ruben, and L. Y. Pao, *A model-free approach to wind farm control using game theoretic methods*, IEEE Trans. Control Syst. Technol., 21 (2013), pp. 1207–1214.
[32] A. Mokhtari, S. Shahrampour, A. Jadbabaie, and A. Ribeiro, *Online optimization in dynamic environments: Improved regret rates for strongly convex problems*, in Proceedings of the 55th IEEE Conference on Decision and Control, IEEE, 2016, pp. 7195–7201.
[33] A. Y. Popkov, *Gradient methods for nonstationary unconstrained optimization problems*, Autom. Remote Control, 66 (2005), pp. 883–891.
[34] S. Rahili and W. Ren, *Distributed continuous-time convex optimization with time-varying cost functions*, IEEE Trans. Automat. Control, 62 (2017), pp. 1590–1605.
[35] B. Shi, S. S. Du, W. Su, and M. I. Jordan, *Acceleration via symplectic discretization of high-resolution differential equations*, in Advances in Neural Information Processing Systems 32, Curran Associates, Red Hook, NY, 2019.
[36] A. Simonetto, *Time-varying Convex Optimization via Time-varying Averaged Operators*, preprint, https://arxiv.org/abs/1704.07338, 2017.
[37] A. Simonetto, E. Dall'Anese, S. Paternain, G. Leus, and G. B. Giannakis, *Time-varying convex optimization: Time-structured algorithms and applications*, Proc. IEEE, 108 (2020), pp. 2032–2048.
[38] A. Simonetto and G. Leus, *Double smoothing for time-varying distributed multiuser optimization*, in Proceedings of the 2014 IEEE Global Conference on Signal and Information Processing, IEEE, 2014, pp. 852–856.
[39] W. Su, S. Boyd, and E. Candes, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, in Advances in Neural Information Processing Systems 27, Curran Associates, Red Hook, NY, 2014.
[40] H. Sun and M. Hong, *Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms*, IEEE Trans. Signal Process., 67 (2019), pp. 5912–5928.
[41] Y. Tang, E. Dall'Anese, A. Bernstein, and S. H. Low, *A feedback-based regularized primal-dual gradient method for time-varying nonconvex optimization*, in Proceedings of the 57th IEEE Conference on Decision and Control, IEEE, 2018, pp. 3244–3250.
[42] Y. Tang, K. Dvijotham, and S. Low, *Real-time optimal power flow*, IEEE Trans. Smart Grid, 8 (2017), pp. 2963–2973.
[43] Y. Tang, G. Qu, and N. Li, *Semi-global exponential stability of augmented primal–dual gradient dynamics for constrained convex optimization*, Systems Control Lett., 144 (2020), 104754.
[44] M. Vaquero and J. Cortés, *Distributed augmentation-regularization for robust online convex optimization*, IFAC-PapersOnLine, 51 (2018), pp. 230–235.
[45] G. Wachsmuth, *On LICQ and the uniqueness of Lagrange multipliers*, Oper. Res. Lett., 41 (2013), pp. 78–80.
[46] D. Willett and J. S. W. Wong, *On the discrete analogues of some generalizations of Gronwall's inequality*, Monatsh. Math., 69 (1965), pp. 362–367.
[47] V. M. Zavala and M. Anitescu, *Real-time nonlinear optimization as a generalized equation*, SIAM J. Control Optim., 48 (2010), pp. 5444–5467, https://doi.org/10.1137/090762634.
[48] Y. Zhang, E. Dall'Anese, and M. Hong, *Online proximal-ADMM for time-varying constrained convex optimization*, IEEE Trans. Signal Inform. Process. Netw., 7 (2021), pp. 144–155.