An Interconnected Systems Approach to Convergence Analysis of Discrete-Time Primal-Dual Algorithms

Spencer Kelly and John W. Simpson-Porco

Abstract— **We study the geometric convergence rate of discrete-time primal-dual algorithms for solving strongly-convex equality-constrained optimization problems. Our approach separates the primal-dual algorithm into an interconnection of two exponentially stable systems, and a composite Lyapunov approach is used to establish stability of the interconnection and provide new bounds on the geometric rate of convergence. Analogous convergence results are developed for two variations of the primal-dual algorithm: an extrapolated version which accelerates convergence, and an inner-loop version which interpolates between the vanilla primal-dual method and dual ascent. The obtained bounds are compared and contrasted with existing bounds from the literature.**

I. INTRODUCTION

Primal-dual algorithms are a broad set of methods for solving constrained optimization problems, with applications across economics [1], signal processing [2], and machine learning. The analysis of such methods has attracted significant attention in recent years from the control community, particularly in the continuous-time setting, in which multiple distinct Lyapunov constructions have been developed for assessing asymptotic, exponential, and semi-global exponential stability; [3]–[13]. Applications of the algorithms have also appeared in the area of feedback-based online optimization [14], [15], wherein they are implemented as real-time dynamic controllers which process measurements.

Our focus in this work will be on primal-dual algorithms for strongly convex problems with linear equality constraints; see Section II for a review. For such problems, it is known that exponential/geometric convergence is achievable, and a key problem of interest is the non-conservative quantification of convergence rates. We will here focus exclusively on algorithms in the discrete-time setting, in which the selection of algorithm step sizes also becomes a key issue.

To review some related work, in [16], a continuous-time primal-dual algorithm is considered for equality-constrained optimization problems, and an Euler discretization is used to show that an associated discrete-time primal-dual algorithm also converges exponentially. The approach limits itself in requiring that the step sizes for both the primal and dual updates are identical. A similar Euler discretization approach is taken in [17] to translate a continuous-time exponential stability result for a proximal primal-dual method into a discrete-time convergence rate.

The authors are with the Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, ON, M5S 3G4, Canada. Email: spen.kelly@mail.utoronto.ca, jwsimpson@ece.utoronto.ca.

This work was supported by University of Toronto Connaught Fund and by NSERC Discovery Grant RGPIN-2017-04008.

Building upon the robust control approach to optimization algorithm analysis in [18], LMI-based methods have recently been used to quantify the convergence rates of primal-dual algorithms in [19]. In [20], a continuous-time primal-dual algorithm is considered in which no rank assumptions are placed on the constraint matrix, and tight bounds are provided on the achievable convergence rates. However, no corresponding rates for discrete-time algorithms are provided, and the frequency-domain analysis does not lead to an explicit parametric Lyapunov construction. Our objective here will be the parametric construction of Lyapunov functions, and in this respect the closest references are [21] and [22], which both study discrete-time primal-dual methods and provide step-size selections and convergence rates; these will be reviewed in detail in Section II-D, and our Lyapunov constructions will be similar to those in [22].

Contributions: Our work here contains three contributions. We first revisit the analysis of [22], providing an interpretation of the analysis method as a loop transformation which separates the dynamics into the interconnection of two stable systems. We improve upon the previous analysis in [22], yielding a less conservative step size selection and improvement in the guaranteed rate; our rate is the best that can be achieved with this particular Lyapunov construction. Second, we extend this analysis to a new family of algorithms which incorporate extrapolation in the primal variable, quantifying how the tuning of the extrapolation parameter leads to acceleration. Finally, inspired by [23] and [20] where primal and dual updates run at different rates, we extend still further the previous analysis technique to study an inner-loop primal dual method, wherein N primal steps are taken in between each dual update. Our analysis quantifies how this inner-loop method smoothly interpolates between the rate of the original primal-dual method at $N = 1$ and the rate of gradient ascent applied to the dual problem as $N \to \infty$.

II. BACKGROUND AND PROBLEM SETUP

A. Preliminaries from Convex Analysis

To begin we recall some basic notions from convex analysis; see, e.g. [24]. A continuously differentiable mapping $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if its gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ satisfies $(\nabla f(x) - \nabla f(y))^{\mathsf{T}}(x - y) \geq 0$ for all $x, y \in \mathbb{R}^n$, and for $m > 0$ is *m-strongly convex* if

$$(\nabla f(x) - \nabla f(y))^{\mathsf{T}}(x - y) \geq m\|x - y\|_2^2$$

for all $x, y \in \mathbb{R}^n$. Similarly, for $L > 0$ we say f is *L-smooth* if ∇f is globally L-Lipschitz continuous, i.e., if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

for all $x, y \in \mathbb{R}^n$. The set of all continuously differentiable m-strongly convex and L-smooth functions will be denoted by $S(m, L)$, and we let $\kappa(f) = L/m \geq 1$ denote the *condition ratio* of f. The *convex conjugate* of f is the function f^* defined by

$$f^*(z) := \sup_{x\in\mathbb{R}^n} \left[x^\mathsf{T} z - f(x)\right].$$

It is a standard result that if $f \in S(m, L)$, then $f^* : \mathbb{R}^n \to \mathbb{R}$, f^* is continuously differentiable, and $f^* \in S(\frac{1}{L}, \frac{1}{m})$. Moreover, in this case, ∇f is invertible and $(\nabla f)^{-1} = \nabla f^*$.

B. Linearly-Constrained Convex Optimization

Our focus will be on first-order algorithms for solving the equality-constrained optimization problem

$$\underset{x\in\mathbb{R}}{\text{minimize}}\ f(x) \quad \text{subject to} \quad Ax = b, \tag{1}$$

where $f \in S(m, L)$ is the objective function to be minimized over the constraint set $\mathsf{C} = \{x \in \mathbb{R}^n \mid Ax = b\}$ with $A \in \mathbb{R}^{r\times n}$ and $b \in \mathbb{R}^r$. With little loss of generality, we assume throughout that A has full row rank, with minimum and maximum singular values $\sigma_{\min}(A) > 0$ and $\sigma_{\max}(A)$, and we let $\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \geq 1$; see [13] for a recent continuous-time study where A is not full row rank.

Under the previous standing assumptions, (1) is feasible and possesses a unique optimal solution $x_\star$. Primal solution techniques such as projected gradient descent

$$x^{k+1} = \mathsf{Proj}_\mathsf{C}(x^k - \alpha\nabla f(x^k)), \qquad \alpha > 0,$$

could be applied to solve (1), but computation of the projection at each step will be expensive in large problems. Instead, methods based on the dual problem can considered. The dual problem associated with (1) can be determined by introducing the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ given by

$$L(x, \lambda) := f(x) + \lambda^\mathsf{T}(Ax - b), \tag{2}$$

where $\lambda \in \mathbb{R}^r$ is the dual variable. Under our assumptions L has a unique saddle point $(x_\star, \lambda_\star)$ where $\lambda_\star$ is the unique optimal solution of the dual problem [25, Theorem 4.1.2]

$$\underset{\lambda\in\mathbb{R}^r}{\text{maximize}} \quad f^*(-A^\mathsf{T}\lambda) - \lambda^\mathsf{T} b \tag{3}$$

with $x_\star$ recoverable as $x_\star = \operatorname{argmin}_x L(x, \lambda_\star) = \nabla f^*(-A^\mathsf{T}\lambda_\star)$. Applying gradient ascent to (3) one obtains

$$\lambda^{k+1} = \lambda^k + \beta(A\nabla f^*(-A^\mathsf{T}\lambda^k) - b), \qquad \beta > 0,$$

which can be expressed as the two step method

$$\tilde{x} = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left[f(x) + (Ax)^\mathsf{T}\lambda^k\right] = \nabla f^*(-A^\mathsf{T}\lambda^k) \tag{4a}$$

$$\lambda^{k+1} = \lambda^k + \beta(A\tilde{x} - b). \tag{4b}$$

Each iteration of (4) requires the solution of the sub-optimization problem (4a), which is again not appealing in large-scale applications.

C. Primal-Dual Algorithms

Primal-dual algorithms iteratively update both the primal and dual variables, with each iteration of the algorithm being considerably simpler than an iteration of the projected gradient descent or dual gradient ascent described above. The most basic such method would be

$$x^{k+1} = x^k - \alpha\left(\nabla f(x^k) + A^\mathsf{T}\lambda^k\right) \tag{5a}$$

$$\lambda^{k+1} = \lambda^k + \beta\left(Ax^k - b\right), \tag{5b}$$

where $\alpha > 0$ and $\beta > 0$ are step sizes. The iteration (5) can be interpreted as simultaneous descent/ascent on the Lagrangian (2) in search of the saddle point, or alternatively, (5a) can be viewed as a crude approximation of (4a). Note that each step of (5) requires only evaluation of ∇f and the evaluation of matrix-vector products involving A and A^T.

The vanilla primal-dual algorithm (5) is but one option. As a generalization, we will also consider in this work the "extrapolated" primal-dual (EPD) algorithm

$$x^{k+1} = x^k - \alpha\left(\nabla f(x^k) + A^\mathsf{T}\lambda^k\right) \tag{6a}$$

$$\tilde{x} = x^k + \tau\left(x^{k+1} - x^k\right) \tag{6b}$$

$$\lambda^{k+1} = \lambda^k + \beta\left(A\tilde{x} - b\right), \tag{6c}$$

where $\tau \in [0, 1]$, $\alpha > 0$, and $\beta > 0$. The parameter τ modifies the value of x used in the dual update (6c), extrapolating from x^k towards x^{k+1}; indeed, (6) reduces to (5) when $\tau = 0$. Our subsequent analysis will first begin with (5) to motivate our approach, and then will be extended to (6).

D. Analysis Methods and Known Rates

Here we review some analysis methods and associated convergence rates from the literature. If $\xi^{k+1} = F(\xi^k)$ denotes a discrete-time system with F continuous and possessing a unique equilibrium $\xi_\star$, the sequence of iterates $\{\xi^k\}_{k\in\mathbb{Z}_{\geq 0}}$ is said to converge to $\xi_\star$ *geometrically with rate* $\rho \in [0, 1)$ if there exists $C > 0$ such that $\|\xi^k - \xi_\star\|_2 \leq C\rho^k\|\xi^0 - \xi_\star\|_2$ for all ξ^0 and all $k \geq 0$. For instance, it is a standard result that if $f \in S(m, L)$, the unconstrained gradient method $x^{k+1} = x^k - \alpha\nabla f(x^k)$ with $\alpha = \frac{2}{m+L}$ converges geometrically with rate $\rho_\mathrm{g} = \frac{\kappa(f)-1}{\kappa(f)+1}$ [18]. Using the properties of the convex conjugate from Section II-A, one may establish that the dual objective in (3) belongs to $S(\bar{m}, \bar{L})$, where we use the short form notation

$$\bar{L} := \frac{\sigma_{\max}(A)^2}{m}, \quad \bar{m} := \frac{\sigma_{\min}(A)^2}{L}, \quad \bar{\kappa} := \frac{\bar{L}}{\bar{m}}. \tag{7}$$

Hence, the dual ascent method (4) with step size $\beta = \frac{2}{\bar{m}+\bar{L}}$ converges geometrically with rate $\rho_\mathrm{d} = \frac{\bar{\kappa}-1}{\bar{\kappa}+1}$.

Returning to the primal-dual methods (5) and (6), we highlight two bounds from the literature on convergence rates. In [22], an analysis technique for (5) based on "ghost sequences" is developed. The method uses the Lyapunov function

$$V(x^k, \lambda^k) = ||x^k - \nabla f^*(-A^\top\lambda^k)||_2 + \omega||\lambda^k - \lambda_\star||_2$$

with $\omega = \frac{L}{m^2}\frac{\sigma_{\max}^3}{\sigma_{\min}^2}$, leading to the step size selections

$$\alpha = \frac{2}{m+L}, \quad \beta = \frac{m}{(m+L)\left(\left(\frac{\sigma_{\max}}{m}\right) + \omega\sigma_{\max}\right)}, \tag{8}$$

and a guaranteed geometric convergence rate of

$$\rho = 1 - \frac{1}{12\kappa(f)^3\kappa(A)^4}. \tag{9}$$

In [21], the authors analyze (6) with $\tau = 1$. The approach uses the quadratic Lyapunov function

$$V(x^k, \lambda^k) = (1 - \alpha\beta\sigma_{\max}^2)||x^k - x^*||_2^2 + \frac{\alpha}{\beta}||\lambda^k - \lambda^*||_2^2$$

with step-sizes satisfying

$$\alpha \leq \frac{1}{L}, \quad \beta \leq \frac{m}{\sigma_{\max}^2(A)}.$$

The approach provides a convergence rate

$$\rho = \max\{1 - \alpha m(1 - \alpha L), 1 - \alpha\beta\sigma_{\min}^2(A)\}.$$

Optimizing the step sizes to achieve the best possible rate in this bound yields

$$\rho^2 = \begin{cases} 1 - \frac{1}{2\kappa(f)}, & \text{if } \kappa(A) \leq \sqrt{2} \\ 1 - \frac{1}{\kappa(f)}\left(\frac{1}{\kappa(A)^2} - \frac{1}{\kappa(A)^4}\right), & \text{otherwise.} \end{cases} \tag{10}$$

These rates will be used as points of comparison for our subsequent results.

III. Improved Convergence Rates via Composite Lyapunov Analysis

One challenge in establishing convergence of (5) or (6) is the lack of internal stability in the dual update when isolated; stability of (5) relies on the stability of the primal variable being "passed through" to the dual update. Our approach in essence will be to perform a loop transformation, separating the dynamics into two exponentially stable subsystems, and then constructing a composite Lyapunov function which establishes exponentially stability of the interconnection. We begin with (5), where the Lyapunov construction is clearer, before proceeding to analyze (6).

A. Analysis of Simultaneous Primal-Dual

Consider the simultaneous primal-dual (SPD) method (5). If one interprets the $A^\mathsf{T}\lambda^k$ term in the primal update (5a) as an external input, then (5a) can be rewritten as a gradient descent algorithm with an external input, given by

$$\begin{aligned} x^{k+1} &= x^k - \alpha\nabla f(x^k) - \alpha u^k \\ u^k &= A^\mathsf{T}\lambda^k. \end{aligned}$$

Next, we interpret (5b) as a perturbation of the *ideal* dual update given by the dual ascent method (4). Adding and subtracting, (5b) can be expressed as

$$\begin{aligned} \lambda^{k+1} &= \lambda^k + \beta\left(A\nabla f^*(-A^\mathsf{T}\lambda^k) - b\right) + \beta A y^k \\ y^k &= x^k - \nabla f^*(-A^\mathsf{T}\lambda^k). \end{aligned}$$

Combining the above ideas, we obtain an alternative representation of the dynamics (5) as the feedback interconnection $\mathcal{F}(\Sigma_1, \Sigma_2)$ of the system Σ_1 defined by

$$\Sigma_1 : \quad \begin{cases} x^{k+1} = x^k - \alpha\nabla f(x^k) - \alpha u^k \\ \quad y^k = x^k - \nabla f^*(-u^k) \end{cases}$$

with the system Σ_2 defined by

$$\Sigma_2 : \quad \begin{cases} \lambda^{k+1} = \lambda^k + \beta\left(A\nabla f(-A^\mathsf{T}\lambda^k) - b\right) + \beta A y^k \\ \quad u^k = A^\mathsf{T}\lambda^k. \end{cases}$$

Note that the output y^k captures the difference between the value x^k used in (5a) and the *ideal* primal value produced by dual ascent in (4). In contrast, Σ_2 is the dual ascent (4), but with an input perturbation. Thus, one may interpret the system Σ_1 as a perturbation to the "ideal" system Σ_2. The following standard result will be used in our analysis.

*Lemma 3.1 (***Contraction of Gradient Descent [22]):** Let $f \in S(m, L)$ and for $\alpha \in \mathbb{R}$ define $F(z) = z - \alpha\nabla f(z)$. If $0 \leq \alpha \leq \frac{2}{m+L}$, then for all $z_1, z_2 \in \mathbb{R}^n$ we have

$$\|F(z_2) - F(z_1)\|_2 \leq (1 - \alpha m)\|z_2 - z_1\|_2.$$

We can now state the first result.

*Theorem 3.2 (***Simultaneous PD Method):** Consider the simultaneous primal-dual algorithm (5) for solving (1) under all previous assumptions, and define the constants

$$\theta_1 := \frac{\sqrt{\bar\kappa(\bar\kappa + 1)}}{\bar\kappa + \sqrt{\bar\kappa(\bar\kappa + 1)}}, \quad \theta_2 := 1 + \bar\kappa + \sqrt{\bar\kappa(\bar\kappa + 1)}$$

where $\bar\kappa$ is as in (7). If the step sizes (α, β) are chosen as

$$\alpha = \frac{2}{m+L}, \quad \beta = \frac{1}{\kappa + 1} \cdot \frac{2}{\theta_1\bar m + \theta_2\bar L}$$

then the iterates of (5) converge geometrically to the unique optimal solution $(x_\star, \lambda_\star)$ of (1) with rate

$$\rho = \rho_{\mathrm{g}} + (1 - \rho_{\mathrm{g}})\frac{\theta_2\bar\kappa}{\theta_1 + \theta_2\bar\kappa}, \text{ where } \rho_{\mathrm{g}} = \frac{\kappa - 1}{\kappa + 1}. \tag{11}$$

Theorem 3.2 provides a modest improvement in guaranteed rate over the analysis in [22]; we defer to the simulation section for a quantitative comparison. The formula for the dual update step size illustrates that the dual steps must be significantly more conservative compared to the optimal tuning $\beta = \frac{2}{\bar m + \bar L}$ for the dual ascent method (4), even when the primal step size $\alpha = \frac{2}{m+L}$ is chosen optimally.

Proof: The system (5) is equivalent to the interconnection of the systems Σ_1 and Σ_2 defined above Lemma 3.1. We will construct a Lyapunov function for the closed-loop system. First, define $V_1(y^k) = \|y^k\|_2$, and compute along closed-loop trajectories that

$$\begin{aligned}V_1(y^{k+1}) &= \|x^{k+1} - \nabla f^*(-u^{k+1})\|_2 \\ &\leq \|x^{k+1} - \nabla f^*(-u^k)\|_2 \\ &\quad + \|\nabla f^*(-u^{k+1}) - \nabla f^*(-u^k)\|_2 \\ &\leq \|x^k - \nabla f^*(-u^k) - \alpha\left(\nabla f(x^k) + u^k\right)\|_2 \\ &\quad + \tfrac{\sigma_{\max}}{m}\|\lambda^{k+1} - \lambda^k\|_2 \\ &= \|x^k - \alpha\nabla f(x^k) - \nabla f^*(-u^k) - \alpha u^k\|_2 \\ &\quad + \tfrac{\sigma_{\max}}{m}\|\lambda^{k+1} - \lambda^k\|_2\end{aligned} \tag{12}$$

Inserting $u^k = -\nabla f(\nabla f^*(-u^k))$ for u_k in the last term within the first normed quantity, for $\alpha \leq \frac{2}{m+L}$ we may invoke Lemma 3.1 with $F(z) = z - \alpha\nabla f(z)$, $z_2 = x^k$, and $z_1 = \nabla f^*(-u^k)$ to further bound as

$$\begin{aligned}V_1(y^{k+1}) \leq&(1 - \alpha m)\|y^k\|_2 \\ &+ \beta\tfrac{\sigma_{\max}}{m}\|A\nabla f^*(-A^\mathsf{T}\lambda^k) - b + Ay^k\|_2 \\ \leq& \left(1 + \beta\tfrac{\sigma_{\max}^2}{m} - \alpha m\right)\|y^k\|_2 \\ &+ \beta\tfrac{\sigma_{\max}^3}{m^2}\|\lambda^k - \lambda_\star\|_2.\end{aligned} \tag{13}$$

The second inequality holds because $f^* \in S(\frac{1}{L}, \frac{1}{m})$ and A is full row-rank. Next, define $V_2(\lambda) = \|\lambda - \lambda_\star\|_2$, and similarly compute along trajectories that

$$\begin{aligned}V_2(\lambda^{k+1}) &= \|\lambda^k - \lambda_\star + \beta A(\nabla f^*(-A^\mathsf{T}\lambda^k) \\ &\quad - \nabla f^*(-A^\mathsf{T}\lambda_\star)) + \beta Ay^k\|_2 \\ &\leq \|\lambda^k - \lambda_\star + \beta(A\nabla f^*(-A^\mathsf{T}\lambda^k) \\ &\quad - A\nabla f^*(-A^\mathsf{T}\lambda_\star))\|_2 + \beta\sigma_{\max}\|y^k\|_2 \\ &\leq (1 - \beta\tfrac{\sigma_{\min}^2}{L})\|\lambda^k - \lambda_\star\|_2 + \beta\sigma_{\max}\|y^k\|_2\end{aligned} \tag{14}$$

for $\beta \leq \frac{2}{\bar{L}+\bar{m}}$ due to Lemma 3.1. Now let $\omega > 0$, and consider the composite Lyapunov function $V(x, \lambda) = V_1(y) + \omega V_2(\lambda)$, where again $y = x - \nabla f^*(-A^\mathsf{T}\lambda)$. By construction, $V(x_\star, \lambda_\star) = 0$ and V is positive definite with respect to $(x_\star, \lambda_\star)$. Combining the inequalities, tedious but routine algebra shows that

$$V(x^{k+1}, \lambda^{k+1}) \leq c_1 V_1(y^k) + c_2\omega V_2(\lambda^k) \tag{15}$$

where

$$\begin{aligned}c_1(\alpha, \beta, \omega) &:= 1 - \alpha m + \beta\left(\bar{L} + \omega\sigma_{\max}\right) \\ c_2(\beta, \omega) &:= 1 + \beta(\bar{L}\tfrac{\sigma_{\max}}{\omega m} - \bar{m}).\end{aligned}$$

The idea is now to select (α, β, ω) to minimize $\max\{|c_1|, |c_2|\}$. First note that since $\alpha \leq \frac{2}{m+L}$, it holds that $c_1 > 0$, and thus since $\alpha \mapsto c_1(\alpha, \beta, \omega)$ is monotonically decreasing for all positive (β, ω), the optimal selection will always be $\alpha = \frac{2}{m+L}$. Substitution and simplification yields

$$\begin{aligned}c_1(\beta, \omega) &:= \tfrac{\kappa-1}{\kappa+1} + \beta(\bar{L} + \omega\sigma_{\max}) \\ c_2(\beta, \omega) &:= 1 - \beta\bar{m}\left(1 - \bar{\kappa}\tfrac{\sigma_{\max}}{\omega m}\right)\end{aligned}$$

For there to exist $\beta > 0$ such that $|c_2| < 1$, we certainly require that $\omega > \bar{\kappa}\frac{\sigma_{\max}}{m}$, in which case, since $\beta \leq \frac{2}{\bar{m}+\bar{L}}$, we must have that $c_2 > 0$. Now observe that (i) c_1 is an increasing function of both β and ω, and (ii) c_2 is a decreasing function of both β and ω. It follows that $\max\{c_1, c_2\}$ will be minimized when $c_1 = c_2$. Equating the expressions and solving for β in terms of ω, one finds that

$$\beta(\omega) = \frac{2}{\kappa+1}\frac{1}{(\bar{L} + \omega\sigma_{\max}) + \bar{m}\left(1 - \bar{\kappa}\frac{\sigma_{\max}}{\omega m}\right)},$$

and substitution into $c(\omega) := c_1(\beta(\omega), \omega)$ yields

$$c(\omega) = \frac{\kappa-1}{\kappa+1} + \frac{2}{\kappa+1}\frac{(\bar{L} + \omega\sigma_{\max})}{(\bar{L} + \omega\sigma_{\max}) + \bar{m}\left(1 - \bar{\kappa}\frac{\sigma_{\max}}{\omega m}\right)}.$$

Through straightforward analysis, one can now argue that $c(\omega)$ possesses a unique minimum over $\omega \in (\bar{\kappa}\frac{\sigma_{\max}}{m}, \infty)$ at

$$\omega_\star = \tfrac{\sigma_{\max}}{m}(\bar{\kappa} + \sqrt{\bar{\kappa}(\bar{\kappa}+1)}),$$

and direct computation of $\beta(\omega_\star) \leq \frac{2}{\bar{m}+\bar{L}}$ and $\rho := c(\omega_\star)$ lead to the step size and rate values in the theorem statement. The inequality (15) now simplifies to $V(x^{k+1}, \lambda^{k+1}) \leq \rho V(x^k, \lambda^k)$ which establishes global exponential stability at the claimed rate. ■

B. Analysis of Extrapolated Primal-Dual

We now extend our analysis to the extrapolated primal-dual method (6). Our approach is again to separate the dynamics (6) into a new feedback interconnection, and following similar reasoning as in Section III-A, we consider the feedback interconnection of the following two input-output systems

$$\Sigma_1: \quad \begin{cases} x^{k+1} = x^k - \alpha\left(\nabla f(x^k) + u^k\right) \\ \quad y_1^k = x^k - \nabla f^*(-u^k) \\ \quad y_2^k = x^{k+1} - \nabla f^*(-u^k), \end{cases}$$

$$\Sigma_2: \quad \begin{cases} \lambda^{k+1} = \lambda^k + \beta\left(A\nabla f^*(-A^\mathsf{T}\lambda^k) - b\right) \\ \qquad\qquad + \beta A(\tau y_1^k + (1-\tau)y_2^k) \\ \quad u^k = A^\mathsf{T}\lambda^k. \end{cases}$$

The interpretation is that the primal subsystem now produces two "error outputs" y_1 and y_2, and the dual subsystem takes these both as inputs and interpolates between them. The previous analysis generalizes to yield the following result.

*Theorem 3.3 (***Extrapolated PD Method***):* Consider the Extrapolated primal-dual (EPD) algorithm (6) for solving (1) under all previous assumptions, and define the constants

$$\theta_1 := \begin{cases} \frac{\sqrt{\bar{\kappa}(\bar{\kappa}+1)}}{\bar{\kappa}+\sqrt{\bar{\kappa}(\bar{\kappa}+1)}}, & \tau \leq \frac{(m+L)(\bar{\kappa}+\sqrt{\bar{\kappa}(\bar{\kappa}+1)})}{\bar{\kappa}+\sqrt{\bar{\kappa}(\bar{\kappa}+1)}+1} \\ 1, & \text{otherwise} \end{cases}$$

$$\theta_2 := \begin{cases} 1 + \bar{\kappa} + \sqrt{\bar{\kappa}(\bar{\kappa}+1)}, & \tau \leq \frac{(m+L)(\bar{\kappa}+\sqrt{\bar{\kappa}(\bar{\kappa}+1)})}{\bar{\kappa}+\sqrt{\bar{\kappa}(\bar{\kappa}+1)}+1}, \\ 1 - \frac{1-\tau+\tau\rho_g}{\tau\rho_g}, & \text{otherwise} \end{cases}$$

Let $\rho_{\mathrm{g}} := \frac{\kappa-1}{\kappa+1} < 1$. If the step sizes (α, β) are chosen as

$$\alpha = \frac{2}{m+L}, \quad \beta = \frac{1}{\kappa+1}\cdot\frac{2}{\theta_1\bar{m} + (1-\tau+\tau\rho_{\mathrm{g}})\theta_2\bar{L}}$$

then the iterates of (5) converge geometrically to the unique optimal solution $(x_\star, \lambda_\star)$ of (1) with rate

$$\rho = \rho_{\mathrm{g}} + (1-\rho_{\mathrm{g}})\frac{(1-\tau+\tau\rho_{\mathrm{g}})\theta_2\bar{\kappa}}{\theta_1 + (1-\tau+\tau\rho_{\mathrm{g}})\theta_2\bar{\kappa}}. \tag{16}$$

Theorem 3.3 allows the user to guarantee a convergence rate for (6) for *any* desired selection of $\tau \in [0,1]$. When $\tau = 0$, Theorem 3.3 reduces exactly to Theorem 3.2. When $\tau = 1$, the given dual update step size becomes

$$\beta = \frac{1}{\kappa+1}\frac{2}{\theta_1\bar{m}+\rho_{\rm g}\theta_2\bar{L}}.$$

Since $\rho_{\rm g}$ (the optimal convergence rate of pure gradient descent) is less than 1, this results in a more aggressive dual update step size compared to that in Theorem 3.2. The corresponding obtained rate

$$\rho = \rho_{\rm g} + (1-\rho_{\rm g})\frac{\rho_{\rm g}\theta_2\bar{\kappa}}{\theta_1+\rho_{\rm g}\theta_2\bar{\kappa}}$$

is faster as a result. Moreover, the corresponding improvement in convergence rate is monotonic in τ, which provides a clear picture of how extrapolation in (6) accelerates convergence compared to (5).

Proof: The method (6) is equivalent to the interconnection of the systems Σ_1 and Σ_2 defined above Theorem 3.3; we construct a Lyapunov function for the closed-loop system. With $y = (y_1, y_2)$, let $V_1(y) = (1-\tau)\|y_1\|_2 + \tau\|y_2\|_2$. Computations quite similar to those in (13) produce the bound

$$\begin{aligned} V_1(y^{k+1}) &\leq (\rho_{\rm g} + \beta(1-\tau+\rho_{\rm g}\tau)\bar{L})V_1(y^k) \\ &\quad + (1-\tau+\tau\rho_{\rm g})\frac{\sigma_{\max}^3}{m^2}\beta\|\lambda^k-\lambda_\star\|_2, \end{aligned} \tag{17}$$

for $\alpha \leq \frac{2}{m+L}$, where one again invokes Lemma 3.1. Defining $V_2(\lambda) = \|\lambda - \lambda_\star\|_2$, computations similar to those in (14) lead to the bound

$$V_2(\lambda^{k+1}) \leq (1-\beta\bar{m})V_1(\lambda^k) + \beta\sigma_{\max}V_1(y^k), \tag{18}$$

for $\beta \leq 2/(\bar{m}+\bar{L})$. For $\omega > 0$ consider the composite Lyapunov function $V(x,\lambda) = V_1(y) + \omega V_2(\lambda)$. Combining the inequalities (17) and (18) and setting $\alpha = \frac{2}{m+L}$ tedious but routine algebra shows that

$$V(x^{k+1},\lambda^{k+1}) \leq \max\{c_1,c_2\}V(x^k,\lambda^k) \tag{19}$$

where

$$\begin{aligned} c_1(\beta,\omega) &= \rho_{\rm g} + \beta((1-\tau+\rho_{\rm g}\tau)\bar{L} + \omega\sigma_{\max}) \\ c_2(\beta,\omega) &= 1 - \beta\bar{m}\left(1-(1-\tau+\rho_{\rm g}\tau)\bar{\kappa}\frac{\sigma_{\max}}{\omega m}\right). \end{aligned} \tag{20}$$

For there to exist a positive β such that $c_2 < 1$, we must necessarily have that $\omega > \delta\bar{\kappa}\frac{\sigma_{\max}}{m}$, where $\delta := 1-\tau+\rho_{\rm g}\tau$ for short. Note that c_1 is an increasing function of both $\beta > 0$ and $\omega > \delta\bar{\kappa}\frac{\sigma_{\max}}{m}$, while c_2 is a decreasing function of both β and ω for the same range. The maximum of c_1 and c_2 therefore will be minimized only when $c_1 = c_2$; equating and solving for β, we obtain

$$\beta(\omega) = \frac{1-\rho_{\rm g}}{(\delta\bar{L}+\omega\sigma_{\max}) + \bar{m}(1-\delta\bar{\kappa}\frac{\sigma_{\max}}{\omega m})} \tag{21}$$

and the back-substitution $c(\omega) := c_1(\beta(\omega),\omega)$ yields

$$c(\omega) = \rho_{\rm g} + (1-\rho_{\rm g})\frac{(\delta\bar{L}+\omega\sigma_{\max})}{(\delta\bar{L}+\omega\sigma_{\max}) + \bar{m}(1-\delta\bar{\kappa}\frac{\sigma_{\max}}{\omega m})}$$

Direct arguments show that $c(\omega)$ achieves its minimum at

$$\omega_\star = \delta\tfrac{\sigma_{\max}}{m}(\bar{\kappa} + \sqrt{\bar{\kappa}(\bar{\kappa}+1)}).$$

Notice that β must also be constrained such that it is compatible with Lemma 3.1. Specifically, it is required that $\beta(\omega) \leq \frac{2}{\bar{m}+\bar{L}}$, which after some tedious calculation holds if

$$\frac{2\tau\sigma_{\max}}{m+L} \leq \omega.$$

The value of $\omega_\star$ decreases below this boundary when

$$\tau \leq \frac{(m+L)(\bar{\kappa}+\sqrt{\bar{\kappa}(\bar{\kappa}+1)})}{\bar{\kappa}+\sqrt{\bar{\kappa}(\bar{\kappa}+1)}+1}. \tag{22}$$

Thus, by defining θ_1 and θ_2 as piecewise functions, it is ensured that when $\omega_\star$ exceeds the lower limit, the value is capped at the minimum. Direct evaluation of $c(\omega_\star)$ and $\beta(\omega_\star)$ followed by simplification yield the expressions in the theorem statement. ∎

IV. Analysis of an Inner-Loop Primal-Dual Algorithm

As previously discussed, the primal-dual algorithms (5) and (6) can be viewed as approximations of the dual ascent method (3), wherein the computation of the dual gradient — an inner optimization problem — is approximated using a *single* primal gradient step. In this section we consider a generalization of (5) involving an inner loop which takes $N \in \mathbb{Z}_{\geq 1}$ primal gradient steps before updating the dual variable. With the dual update function $\mathscr{D}(x,\lambda) := \lambda + \beta(Ax-b)$, we express this as

$$\begin{aligned} x^{k+1} &= x^k - \alpha(\nabla f(x^k) + A^{\mathsf{T}}\lambda^k) \\ \lambda^{k+1} &= \begin{cases} \mathscr{D}(x^k,\lambda^k), & k \bmod N = 0 \\ \lambda^k, & \text{otherwise.} \end{cases} \end{aligned} \tag{23}$$

One may think of this as updating the dual variable only at steps $\{0, N, 2N, \ldots\}$. When $N = 1$ we recover (5), and as $N \to \infty$, we should intuitively recover the dual ascent method (3). Our result below captures this intuition.

Theorem 4.1 (**Inner-Loop Primal-Dual):** Consider the inner-loop primal-dual algorithm (23) for solving (1) under all previously stated assumptions. Let $\epsilon = \rho_{\rm g}^{N-1}$, set

$$\begin{aligned} a &:= 2\sigma_{\max}(L+m) \\ b &:= (\bar{L}+\bar{m})((1-\epsilon)L + (1+\epsilon)m) - 2(L+m)(\epsilon\bar{L}+\bar{m}) \\ c &:= 2\epsilon\bar{L}\frac{\sigma_{\max}}{m}(L+m)(\bar{m}+\bar{L}), \end{aligned} \tag{24}$$

and define

$$\omega := \max\left\{\epsilon\frac{\sigma_{\max}}{m}\left(\bar{\kappa}+\sqrt{\bar{\kappa}(\bar{\kappa}+1}\right), \frac{b+\sqrt{b^2+4ac}}{2a}\right\}.$$

If the step sizes (α,β) are chosen as

$$\alpha = \tfrac{2}{m+L}, \; \beta = \tfrac{(1-\epsilon)L+(1+\epsilon)m}{(L+m)(\epsilon\bar{L}+\omega\sigma_{\max}+\bar{m}-\epsilon\bar{L}\frac{\sigma_{\max}}{\omega m})} \tag{25}$$

then the iterates of (23) converge geometrically to the unique optimal solution with rate

$$\rho = 1 - \frac{((1-\epsilon)L+(1+\epsilon)m)(\bar{m}-\epsilon\bar{L}\frac{\sigma_{\max}}{\omega m})}{(L+m)(\epsilon\bar{L}+\omega\sigma_{\max}+\bar{m}-\epsilon\bar{L}\frac{\sigma_{\max}}{\omega m})}. \tag{26}$$

While the above expressions are complex to parse, the rates guaranteed in Theorem 4.1 do indeed match with the intuition behind the algorithm. When $N = 1$, $\epsilon = 1$ and some rather involved calculations show that the provided rate reduces to the rate given for the simultaneous primal-dual method in Theorem 3.2. Conversely, as $N \to \infty$, we have $\epsilon \to 0$, $\omega \to \frac{b}{a} = \frac{\bar{L}-\bar{m}}{2\sigma_{\max}}$, and

$$\rho \to 1 - \frac{\bar{m}}{\bar{m} + \frac{\bar{L}-\bar{m}}{2}} = 1 - \frac{2\bar{m}}{\bar{L}+\bar{m}} = \frac{\bar{\kappa}-1}{\bar{\kappa}+1}$$

which is the optimal rate of dual ascent (4). Thus, the results of the theorem interpolate between the simultaneous primal-dual method (5) and the dual ascent method (4). The proof is a slightly more complex version of the previously presented proofs, and is omitted.

V. Numerical Comparisons of Rates

A. Simultaneous Primal-Dual Algorithm

The rate obtained in Theorem 3.2, given by (11), will be compared against (9), the rate achieved in [22]. Two plots will be used. First, the condition number of A, $\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$, will be held constant while the condition ratio of the function f, $\kappa(f) = \frac{L}{m}$, is varied. Figure 1 illustrates the results of holding $\kappa(A) = 1.5$ and varying $\kappa(f)$ between 1 and 20. Figure 2 illustrates the results of holding $\kappa(f) = 1.5$ and varying $\kappa(A)$ between 1 and 20. It is clear the result achieved in Theorem 3.2 improves upon (9). Although the Lyapunov approach was the same, meticulously selecting the parameters allowed for an improvement of the rate.

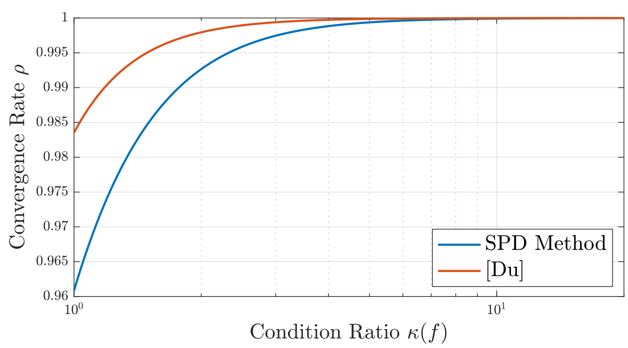

Fig. 1: Convergence rate as κ varies from 1 to 20 while $\kappa(A) = 1.5$

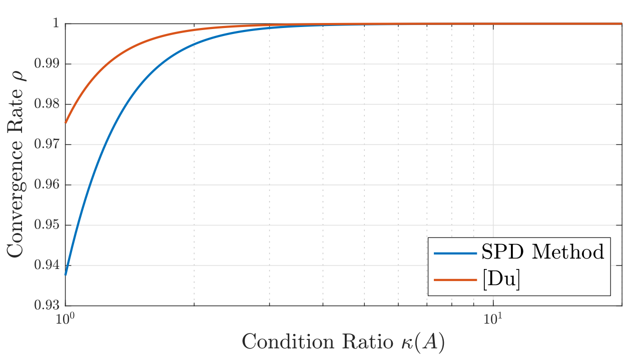

Fig. 2: Convergence rate as $\kappa(A)$ varies from 1 to 20 while $\kappa = 1.5$.

B. Extrapolated Primal-Dual Algorithm

For $\tau = 0$, the rate of the extrapolated primal-dual algorithm (16) matches that of the simultaneous primal-dual algorithm (11), but the former rate improves as τ increases. Figure 3 and 4 illustrate the results of holding $\tau = 1$ and varying $\kappa(f)$ and $\kappa(A)$, respectively. Both approaches are compared against (10), achieved in [21]. Notice for small values of $\kappa(f)$ and $\kappa(A)$, the rate guaranteed in Theorem 3.3 provides generous bounds, surpassing those of (10); however, it lags behind as the values increase. As future work, this suggests that a careful combination of the Lyapunov approach here and that in [21] may be extremely effective.

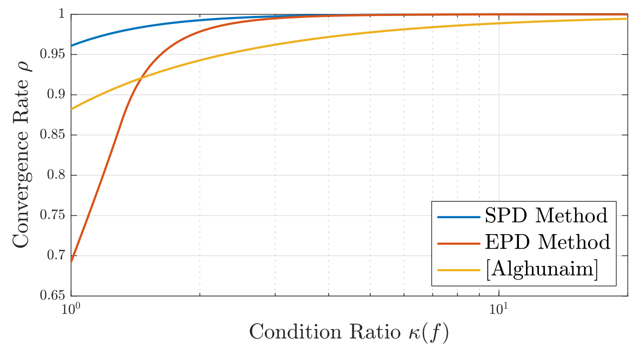

Fig. 3: Convergence rate as κ varies from 1 to 20 while $\kappa(A) = 1.5$.

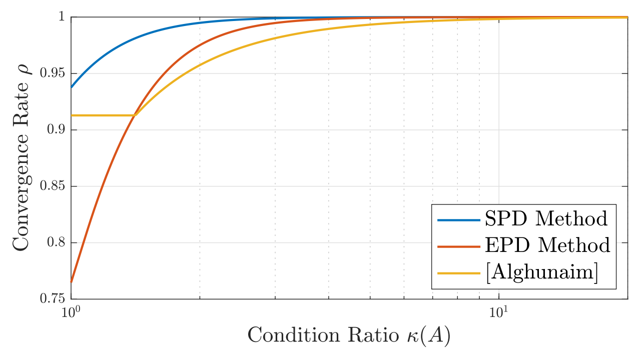

Fig. 4: Convergence rate as $\kappa(A) = 1.5$ varies from 1 to 20

C. Inner-Loop Primal-Dual Algorithm

The rate of convergence of the inner-loop primal-dual method (26) is a function of N. When $N = 1$, the rate equals that of the SPD method, and as N increases, the rate improves and asymptotically approaches the rate of the dual ascent method. Moreover, one may show that when $N = 2$, the inner-loop primal-dual method outperforms the extrapolated primal-dual algorithm rate. Figure 5 and Figure 6 display the rates achieved while setting $N = 20$ and varying $\kappa(f)$ and $\kappa(A)$, respectively. Not surprisingly, the inner loop method performs favourably, as it uses more primal gradient evaluations per step than the other methods. Curiously however, when $\kappa(f)$ increases, the inner-loop falls behind (10); this illustrates the importance of extrapolation, which is absent in the inner-loop method studied here. Increasing N further can improve these results and provide a rate that surpasses (10).

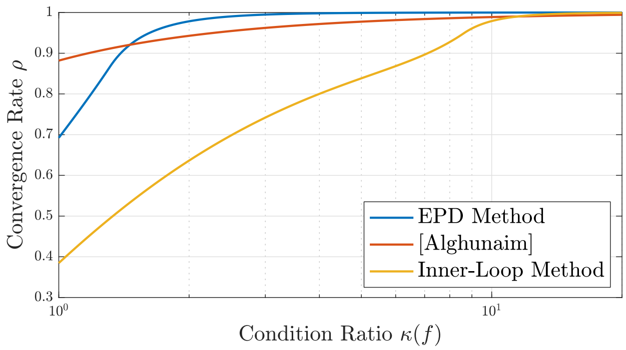

Fig. 5: Convergence rate as $\kappa(f)$ varies from 1 to 20 with $\kappa(A) = 1.5$, $\tau = 1$, $N = 20$

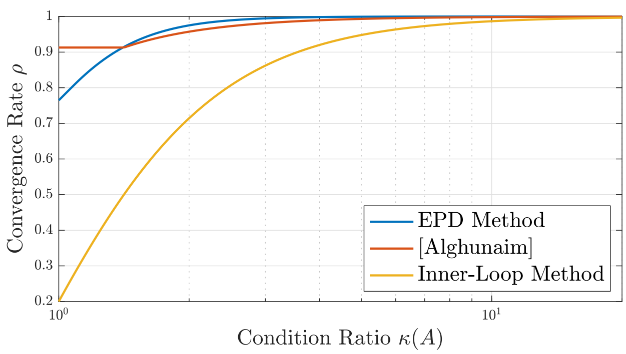

Fig. 6: Convergence rate as $\kappa(A)$ varies from 1 to 20 with $\kappa(f) = 1.5$, $\tau = 1$, $N = 20$

When $\kappa(A)$ is varied, the inner-loop primal-dual algorithm significantly outperforms the two other rates.

VI. Conclusion

We have analyzed discrete-time primal-dual algorithms using an interconnected systems lens, deriving new step size selectio1ns and convergence rates for vanilla, extrapolated, and inner-loop versions of the methods. The proof approach broadly followed and expanded on the approach proposed by [22], but with improved parameter selections and extensions to the extrapolated and inner-loop algorithms. As directions for future work, taking an interconnected systems lens for the analysis of more sophisticated primal-dual algorithms appears to be promising, as this could allow for the systematic study of, e.g., accelerated primal-dual methods, and primal-dual methods involving non-smooth regularizers [6], [15].

References

[1] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in linear and non-linear programming*. Stanford Univ. Press, 1972.

[2] A. Salim, L. Condat, D. Kovalev, and P. Richtarik, "An optimal algorithm for strongly convex minimization under affine constraints," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Eds., vol. 151. PMLR, 28–30 Mar 2022, pp. 4482–4498. [Online]. Available: https://proceedings.mlr.press/v151/salim22a.html

[3] D. Feijer and F. Paganini, "Stability of primal–dual gradient dynamics and applications to network optimization," *Automatica*, vol. 46, no. 12, pp. 1974–1981, 2010.

[4] A. Cherukuri, E. Mallada, and J. Cortés, "Asymptotic convergence of constrained primal–dual dynamics," *IFAC Syst & Control L*, vol. 87, pp. 10 – 15, 2016.

[5] J. W. Simpson-Porco, "Input/output analysis of primal-dual gradient algorithms," in *Allerton Conf on Comm, Ctrl & Comp*, Monticello, IL, USA, Sep. 2016, pp. 219–224.

[6] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented lagrangian method for nonsmooth composite optimization," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2861–2868, 2019.

[7] A. Cherukuri, E. Mallada, S. Low, and J. Cortés, "The role of convexity on saddle-point dynamics: Lyapunov function and robustness," *IEEE Trans. Autom. Control*, vol. PP, no. 99, pp. 1–1, 2017.

[8] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented lagrangian method for nonsmooth composite optimization," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2861–2868, 2019.

[9] D. Ding and M. R. Jovanović, "Global exponential stability of primal-dual gradient flow dynamics based on the proximal augmented lagrangian," in *Proc. ACC*, Philadelphia, PA, USA, Jul. 2019, pp. 3414–3419.

[10] X. Chen and N. Li, "Exponential stability of primal-dual gradient dynamics with non-strong convexity," in *Proc. ACC*, Denver, CO, USA, Jul. 2020, pp. 1612–1618.

[11] Y. Tang, G. Qu, and N. Li, "Semi-global exponential stability of augmented primal–dual gradient dynamics for constrained convex optimization," *IFAC Syst & Control L*, vol. 144, p. 104754, 2020.

[12] I. K. Ozaslan and M. R. Jovanović, "Exponential convergence of primal-dual dynamics for multi-block problems under local error bound condition," in *Proc. IEEE CDC*, 2022, pp. 7579–7584.

[13] ——, "On the global exponential stability of primal-dual dynamics for convex problems with linear equality constraints," in *Proc. ACC*, 2023, pp. 210–215.

[14] A. Bernstein, E. Dall'Anese, and A. Simonetto, "Online primal-dual methods with measurement feedback for time-varying convex optimization," *IEEE Trans. Signal Proc.*, vol. 67, no. 8, pp. 1978–1991, 2019.

[15] M. Colombino, E. Dall'Anese, and A. Bernstein, "Online optimization as a feedback controller: Stability and tracking," *IEEE Trans. Control Net. Syst.*, vol. 7, no. 1, p. 422–432, 2020.

[16] G. Qu and N. Li, "On the exponential stability of primal-dual gradient dynamics," *IEEE Control Syst. Let.*, vol. 3, no. 1, pp. 43–48, 2019.

[17] D. Ding and M. R. Jovanović, "Global exponential stability of primal-dual gradient flow dynamics based on the proximal augmented lagrangian: A lyapunov-based approach," in *Proc. IEEE CDC*, Jeju, Korea, Dec. 2020, pp. 4836–4841.

[18] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2016. [Online]. Available: https://doi.org/10.1137/15M1009597

[19] B. Van Scoy, J. W. Simpson-Porco, and L. Lessard, "Automated Lyapunov analysis of primal-dual optimization algorithms: An interpolation approach," in *Proc. IEEE CDC*, Singapore, Dec. 2023, accepted to IEEE CDC on July 11th, 2023.

[20] I. K. Ozaslan and M. R. Jovanović, "Tight lower bounds on the convergence rate of primal-dual dynamics for equality constrained convex problems," in *Proceedings of the 62nd IEEE Conference on Decision and Control*, Singapore, 2023, pp. 7312–7317.

[21] S. A. Alghunaim and A. H. Sayed, "Linear convergence of primal–dual gradient methods and their performance in distributed optimization," *Automatica*, vol. 117, p. 109003, 2020.

[22] S. S. Du and W. Hu, "Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 196–205. [Online]. Available: https://proceedings.mlr.press/v89/du19b.html

[23] M. T. Hale, A. Nedić, and M. Egerstedt, "Asynchronous multiagent primal-dual optimization," *IEEE Trans. Autom. Control*, vol. 62, no. 9, pp. 4421–4435, 2017.

[24] E. K. Ryu and S. Boyd, "A primer on monotone operator methods," *Applied and Computational Mathematics*, vol. 15, no. 1, pp. 3–43, 2016.

[25] A. Auslender and M. Teboulle, *Asymptotic cones and functions in optimization and variational inequalities*. Springer, 2011.