On the Exponential Stability of Primal-Dual Gradient Dynamics*

Guannan Qu[1] and Na Li[1]

***Abstract*— Continuous time primal-dual gradient dynamics that find a saddle point of a Lagrangian of an optimization problem have been widely used in systems and control. While the global asymptotic stability of such dynamics has been well-studied, it is less studied whether they are globally exponentially stable. In this paper, we study the primal-dual gradient dynamics for convex optimization with strongly-convex and smooth objectives and affine equality or inequality constraints, and prove global exponential stability for such dynamics. Bounds on decaying rates are provided.**

I. INTRODUCTION

This paper considers the following constrained optimization problem

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \\ \text{s.t.}\quad & A_1\mathbf{x} = \mathbf{b}_1 \\ & A_2\mathbf{x} \le \mathbf{b}_2 \end{aligned} \tag{1}$$

where $A_1 \in \mathbb{R}^{m_1\times n}, A_2 \in \mathbb{R}^{m_2\times n}$ and $\mathbf{b}_1 \in \mathbb{R}^{m_1}, \mathbf{b}_2 \in \mathbb{R}^{m_2}$ and $f(\mathbf{x})$ is a strongly convex and smooth function. Let $L(\mathbf{x},\boldsymbol{\lambda})$ be the Lagrangian (or Augmented Lagrangian) associated with Problem (1). The focus of this paper is the following primal-dual gradient dynamics, also known as saddle-point dynamics, associated with the Lagrangian $L(\mathbf{x},\boldsymbol{\lambda})$,

$$\dot{\mathbf{x}} = -\eta_1\nabla_{\mathbf{x}}L(\mathbf{x},\boldsymbol{\lambda}) \tag{2a}$$

$$\dot{\boldsymbol{\lambda}} = \eta_2\nabla_{\boldsymbol{\lambda}}L(\mathbf{x},\boldsymbol{\lambda}) \tag{2b}$$

where $\eta_1, \eta_2 > 0$ are time constants.

Primal-Dual Gradient Dynamics (PDGD), also known as saddle-point dynamics, were first introduced in [1], [2]. They have been widely used in engineering and control systems, for example in power grid [3], [4], wireless communication [5], [6], network and distributed optimization [7], [8], game theory [9], etc. Despite its wide applications, general studies on PDGD [1], [2], [8], [10]–[22] have mostly focused on its asymptotic stability (or convergence), with few studying its global exponential stability. It is known that the gradient dynamics for the unconstrained version of (1) achieves global exponential stability when f is strongly convex and smooth. It is natural to raise the question whether in the constrained case, PDGD can also achieve global exponential stability.

Global exponential stability is a desirable property in practice. Firstly, in control systems especially those in critical infrastructure like the power grid, it is desirable to have strong stability guarantees. Secondly, when using PDGD as computational tools for constrained optimization, discretization is essential for implementation. The global exponential stability ensures that the simple explicit Euler discretization has a geometric convergence rate when the discretization step size is sufficiently small [23], [24]. This is an appealing property for discrete-time optimization methods.

Contribution of this paper. In this paper, we prove the global exponential stability of PDGD (2) under some regularity conditions on problem (1) and we also give bounds on the decaying rates (Theorem 1 and 2). Our proof relies on a quadratic Lyapunov function that has non-zero off-diagonal terms, which is different from the (block-)diagonal quadratic Lyapunov functions that are commonly used in the literature [6], [20] and are known being unable to certify global exponential stability [6, Lemma 3]. We also highlight that when handling inequality constraints, we use a variant of the PDGD based on Augmented Lagrangian [25] and is projection free. This is different from the projection-based PDGD studied in [18], [20], which is discontinuous (see Footnote 2 for more discussions). Our variant of PDGD guarantees that the multipliers stay nonnegative without using projection, and avoids the discontinuity problem caused by projection [18], [20] (see Remark 1 for more discussions).

A. Related Work

There have been many efforts in studying the stability of PDGD as well as its discrete time version. An incomplete list includes [1], [2], [8], [10]–[22]. For instance, [17] studies the subgradient saddle point algorithm and proves its convergence to an approximate saddle point with rate $O(\frac{1}{t})$; [18] uses LaSalle invariance principle to prove global asymptotic stability of PDGD; [22] studies the global asymptotic stability of the saddle-point dynamics associated with general saddle functions; and [20] proves global asymptotic stability of PDGD with projection, which is extended in [21] by using a weaker assumption and proving input-to-state stability.

Our work is closely related to a recent paper [8] which studies saddle-point-like dynamics and proves global exponential stability when applying such dynamics to equality constrained convex optimization problems. The difference between [8] and our work is that for the equality constrained case, [8] considers a different Lagrangian from ours. Further, the result in [8] cannot be directly generalized to inequality constrained case [8, Remark 3.9].

Our work is also related to the vast literature on spectral bounds on saddle matrices [26], [27]. Such bounds, when combined with Ostrowski Theorem [28, 10.1.4], can lead to *local* exponential stability results of PDGD [25, Sec 4.4.1] [29, Prop. 4.4.1] as opposed to *global* exponential stability which is the focus of this paper.

It recently came to our attention that [30] studies a class of dynamics, a special case of which turns out to

*This work is supported by NSF 1608509, NSF CAREER 1553407, and ARPA-E NODES.

[1] Guannan Qu and Na Li are with John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge MA 02138, USA. Emails: gqu@g.harvard.edu, nali@seas.harvard.edu

be similar to our PDGD for affine inequality constraints. [30] also proves global exponential stability. Their proof uses frequency domain analysis, which is different from our time-domain analysis. It remains interesting to investigate the connection between the methods of [30] and our work.

Notations. Throughout the paper, scalars will be small letters, vectors will be bold small letters and matrices will be capital letters. Notation $\|\cdot\|$ represents Euclidean norm for vectors, and spectrum norm for matrices. For any symmetric matrix P_1, P_2 of the same dimension, $P_1 \succeq P_2$ means $P_1 - P_2$ is positive semi-definite.

II. Algorithms and Main Results

In this section we describe our PDGD for solving Problem (1) and present stability results. Throughout this paper, we use the following assumption of f:

Assumption 1. *Function f is twice continuously differentiable, μ-strongly convex and ℓ-smooth, i.e. for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$\mu\|\mathbf{x}-\mathbf{y}\|^2 \leq \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \leq \ell\|\mathbf{x}-\mathbf{y}\|^2 \quad (3)$$

To streamline exposition, we will present the equality constrained case and the inequality constrained case separately. Integrating them will give PDGD with global exponential stability for Problem (1). Without causing any confusion, notations will be double-used in the two cases.

A. Equality Constrained Case

We first consider the equality constrained case,

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} \end{aligned} \quad (4)$$

Here we remove the subscript for A and $\mathbf{b}$ in Problem (1) for notational simplicity. Problem (4) has the Lagrangian,

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T(A\mathbf{x}-\mathbf{b}) \quad (5)$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$ is the Lagrangian multiplier. The PDGD is,

$$\dot{\mathbf{x}} = -\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = -\nabla f(\mathbf{x}) - A^T\boldsymbol{\lambda} \quad (6a)$$

$$\dot{\boldsymbol{\lambda}} = \eta \nabla_{\lambda} L(\mathbf{x}, \boldsymbol{\lambda}) = \eta(A\mathbf{x}-\mathbf{b}) \quad (6b)$$

where without loss of generality, we have fixed the time constant of the primal part to be 1. We make the following assumption on A, which is the linear independence constraint qualification for (4).

Assumption 2. *We assume that matrix A is full row rank and $\kappa_1 I \preceq AA^T \preceq \kappa_2 I$ for some $\kappa_1, \kappa_2 > 0$.*

Let $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ be the equilibrium point of (6), which in this case is also the saddle point of L.[1] The following theorem gives the global exponential stability of the PDGD (6).

Theorem 1. *Under Assumption 1 and 2, for $\eta > 0$, define $\tau_{eq} = \min(\frac{\eta\kappa_1}{4\ell}, \frac{\kappa_1\mu}{4\kappa_2})$. Then there exist constants C_1, C_2 that depend on $\eta, \kappa_1, \kappa_2, \mu, \ell$, $\|\mathbf{x}(0)-\mathbf{x}^*\|$, $\|\boldsymbol{\lambda}(0)-\boldsymbol{\lambda}^*\|$, s.t. $\|\mathbf{x}(t)-\mathbf{x}^*\| \leq C_1 e^{-\frac{1}{2}\tau_{eq}t}$ and $\|\boldsymbol{\lambda}(t)-\boldsymbol{\lambda}^*\| \leq C_2 e^{-\frac{1}{2}\tau_{eq}t}$.*

[1] Assumption 1 and 2 guarantee that the saddle point exists and is unique.

B. Inequality Constrained Case

Now we consider the inequality constrained case,

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & A\mathbf{x} \leq \mathbf{b} \end{aligned} \quad (7)$$

where f and A satisfy Assumption 1 and 2. For the inequality constrained case, we use the "Augmented Lagrangian" [25, Sec. 3.1], as opposed to the standard Lagrangian in [18], [20].[2] In details, let $A^T = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m]$, with each $\mathbf{a}_j \in \mathbb{R}^n$, and let $\mathbf{b} = [b_1, \ldots, b_m]^T$. Then we define the augmented Lagrangian,

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{j=1}^m H_\rho(\mathbf{a}_j^T\mathbf{x} - b_j, \lambda_j) \quad (8)$$

where $\rho > 0$ is a free parameter, $H_\rho(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$ is a penalty function on constraint violation, defined as follows

$$\begin{aligned} & H_\rho(\mathbf{a}_j^T\mathbf{x} - b_j, \lambda_j) \\ & = \begin{cases} (\mathbf{a}_j^T\mathbf{x} - b_j)\lambda_j & \\ \quad + \frac{\rho}{2}(\mathbf{a}_j^T\mathbf{x} - b_j)^2 & \text{if } \rho(\mathbf{a}_j^T\mathbf{x} - b_j) + \lambda_j \geq 0 \\ -\frac{1}{2}\frac{\lambda_j^2}{\rho} & \text{if } \rho(\mathbf{a}_j^T\mathbf{x} - b_j) + \lambda_j < 0 \end{cases} \end{aligned}$$

We can then calculate the gradient of H_ρ w.r.t. $\mathbf{x}$ and $\boldsymbol{\lambda}$.

$$\nabla_{\mathbf{x}} H_\rho(\mathbf{a}_j^T\mathbf{x} - b_j, \lambda_j) = \max(\rho(\mathbf{a}_j^T\mathbf{x} - b_j) + \lambda_j, 0)\mathbf{a}_j$$

$$\nabla_{\boldsymbol{\lambda}} H_\rho(\mathbf{a}_j^T\mathbf{x} - b_j, \lambda_j) = \frac{\max(\rho(\mathbf{a}_j^T\mathbf{x} - b_j) + \lambda_j, 0) - \lambda_j}{\rho}\mathbf{e}_j$$

where $\mathbf{e}_j \in \mathbb{R}^m$ is a vector with the j'th entry being 1 and other entries being 0. The primal-dual gradient dynamics for the augmented Lagrangian L is given in (9). We call it as Aug-PDGD (Augmented Primal-Dual Gradient Dynamics).

$$\begin{aligned} \dot{\mathbf{x}} &= -\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = -\nabla f(\mathbf{x}) - \sum_{j=1}^m \nabla_{\mathbf{x}} H_\rho(\mathbf{a}_j^T\mathbf{x} - b_j, \lambda_j) \\ &= -\nabla f(\mathbf{x}) - \sum_{j=1}^m \max(\rho(\mathbf{a}_j^T\mathbf{x} - b_j) + \lambda_j, 0)\mathbf{a}_j \end{aligned} \quad (9a)$$

$$\begin{aligned} \dot{\boldsymbol{\lambda}} &= \eta\nabla_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) = \eta\sum_{j=1}^m \nabla_{\boldsymbol{\lambda}} H_\rho(\mathbf{a}_j^T\mathbf{x} - b_j, \lambda_j) \\ &= \eta\sum_{j=1}^m \frac{\max(\rho(\mathbf{a}_j^T\mathbf{x} - b_j) + \lambda_j, 0) - \lambda_j}{\rho}\mathbf{e}_j \end{aligned} \quad (9b)$$

Remark 1. *It is easy to check that, if $\lambda_j(0) \geq 0$, then (9b) guarantees $\lambda_j(t) \geq 0, \forall t$. This means that the dynamics (9) automatically guarantees $\lambda_j(t)$ will stay nonnegative as long as its initial value is nonnegative, without using projection as is done in [18], [20], thus avoiding discontinuity issues caused by the projection step.*

Since the saddle point of the Augmented Lagrangian (8)

[2] We also considered using the standard Lagrangian and the associated PDGD with projection in [18], [20]. However, the discontinuous projection step therein created difficulties both in theoretic analysis and numerical simulations. Theoretically, the projection step is based on Euclidean norm, and is consistent with the block diagonal Lyapunov function used in [18], [20], but it is not consistent with the Lyapunov function with cross term used in this paper (cf. (14)), which is the key for proving global exponential stability. Therefore we conjecture that the PDGD with projection [18], [20] is *not* exponentially stable. Numerically, when we simulate the PDGD with the discontinuous projection step using MATLAB ODE solvers, we encounter many numerical issues. For these reasons, in this paper we study an alternative projection-free PDGD based on the Augmented Lagrangian.

is the same as that of the standard Lagrangian (see [25, Sec. 3.1] for details), we have the following proposition regarding the equilibrium point $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ of Aug-PDGD. For completeness we include a proof in our online report [31, Appendix-E].

Proposition 1 ([25]). *Under Assumption 1 and 2, Aug-PDGD (9) has a unique equilibrium point $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ and it satisfies the KKT condition of problem (7).*

Aug-PDGD (9) is globally exponentially stable, as stated below.

Theorem 2. *Under Assumption 1 and 2, the Aug-PDGD* (9) *is globally exponentially stable in the sense that, for any $\eta > 0$, $\rho > 0$, there exists constant*

$$\tau_{ineq} = \frac{\eta\kappa_1^2}{40\ell\kappa_2[\max(\frac{\rho\kappa_2}{\mu}, \frac{\ell}{\mu})]^2[\max(\frac{\eta}{\ell\rho}, \frac{\ell}{\mu})]^2}$$

and constants $C_3, C_4 > 0$ which depend on $\eta, \rho, \kappa_1, \kappa_2, \mu, \ell$, $\|\mathbf{x}(0) - \mathbf{x}^*\|$, $\|\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}^*\|$, s.t. $\|\mathbf{x}(t) - \mathbf{x}^*\| \le C_3 e^{-\frac{1}{2}\tau_{ineq}t}$, $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}^*\| \le C_4 e^{-\frac{1}{2}\tau_{ineq}t}$.

Remark 2. *We currently only study the affine inequality constrained case and assume the matrix A satisfies Assumption 2. We conjecture that the results can be extended in two ways. Firstly, for affine inequality constraints, Assumption 2 can be relaxed to the linear independence constraint qualification, i.e. at the optimizer $\mathbf{x}^*$, the submatrix of A associated with the active constraints has full row rank. Secondly, for nonlinear convex constraint $\mathbf{J}(x) \le 0$ where $\mathbf{J} : \mathbb{R}^n \to \mathbb{R}^m$, Assumption 2 can be replaced by the condition $\kappa_1 I \preceq \frac{\partial \mathbf{J}(\mathbf{x})}{\partial \mathbf{x}}(\frac{\partial \mathbf{J}(\mathbf{x})}{\partial \mathbf{x}})^T \preceq \kappa_2 I$ where $\frac{\partial \mathbf{J}(\mathbf{x})}{\partial \mathbf{x}} \in \mathbb{R}^{m \times n}$ is the Jacobian of $\mathbf{J}$ w.r.t. $\mathbf{x}$. We leave these extensions to the future work.*

III. Stability Analysis

In this section, we prove global exponential stability. We also show global exponential stability ensures the geometric convergence rate of the Euler discretization.

A. The Equality Constrained Case, Proof of Theorem 1

We stack $\mathbf{x}$ and $\boldsymbol{\lambda}$ into a larger vector $\mathbf{z} = [\mathbf{x}^T, \boldsymbol{\lambda}^T]^T$ and similarly define $\mathbf{z}^* = [(\mathbf{x}^*)^T, (\boldsymbol{\lambda}^*)^T]^T$. We define quadratic Lyapunov function, $V(\mathbf{z}) = (\mathbf{z} - \mathbf{z}^*)^T P (\mathbf{z} - \mathbf{z}^*)$ with $P \succ 0$ defined by

$$P = \begin{bmatrix} \eta c I & \eta A^T \\ \eta A & cI \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)} \tag{10}$$

where $c = 4\max(\ell, \frac{\eta\kappa_2}{\mu})$.[3] If we can show the following property of $V(\mathbf{z})$ along the trajectory of the dynamics,

$$\frac{d}{dt}V(\mathbf{z}) \le -\tau V(\mathbf{z}) \tag{11}$$

for $\tau = \frac{\eta\kappa_1}{c} = \min(\frac{\eta\kappa_1}{4\ell}, \frac{\kappa_1\mu}{4\kappa_2})$, then we have proved Theorem 1. The rest of the section will be devoted to proving (11). We start with the following auxiliary Lemma, which can be proved by using mean value theorem. A similar lemma can be found in [8, Lem. A.1], and for completeness we include a proof in our online report [31, Appendix-D].

Lemma 1. *Under Assumption 1, for any $\mathbf{x} \in \mathbb{R}^n$, there exists a symmetric matrix $B(\mathbf{x})$ that depends on $\mathbf{x}$, satisfying $\mu I \preceq B(\mathbf{x}) \preceq \ell I$, s.t. $\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*) = B(\mathbf{x})(\mathbf{x} - \mathbf{x}^*)$.*

With Lemma 1, we can rewrite PDGD (6) as,

$$\begin{aligned} \frac{d}{dt}\mathbf{z} &= \begin{bmatrix} -(\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) - \nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)) \\ \eta\nabla_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) - \eta\nabla_{\boldsymbol{\lambda}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \end{bmatrix} \\ &= \begin{bmatrix} -B(\mathbf{x})(\mathbf{x} - \mathbf{x}^*) - A^T(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*) \\ \eta A(\mathbf{x} - \mathbf{x}^*) \end{bmatrix} \\ &= \begin{bmatrix} -B(\mathbf{x}) & -A^T \\ \eta A & 0 \end{bmatrix} (\mathbf{z} - \mathbf{z}^*) := G(\mathbf{z})(\mathbf{z} - \mathbf{z}^*). \end{aligned} \tag{12}$$

Then, $\frac{d}{dt}V(\mathbf{z})$ can be written as

$$\begin{aligned} \frac{d}{dt}V(\mathbf{z}) &= \dot{\mathbf{z}}^T P(\mathbf{z} - \mathbf{z}^*) + (\mathbf{z} - \mathbf{z}^*)^T P\dot{\mathbf{z}} \\ &= (\mathbf{z} - \mathbf{z}^*)^T (G(\mathbf{z})^T P + PG(\mathbf{z}))(\mathbf{z} - \mathbf{z}^*) \end{aligned} \tag{13}$$

Therefore, to prove (11), it is sufficient to prove the following Lemma, whose proof is in Appendix-A.

Lemma 2. *For any $\mathbf{z} \in \mathbb{R}^{n+m}$, we have*

$$G(\mathbf{z})^T P + PG(\mathbf{z}) \preceq -\tau P$$

Lemma 2 and (13) lead to (11), concluding the proof.

B. The Inequality Constrained Case, Proof of Theorem 2

We start by emphasizing the notations in this section is independent from the equality constrained case in Section III-A. We stack $\mathbf{x}$, $\boldsymbol{\lambda}$ into a larger vector $\mathbf{z} = [\mathbf{x}^T, \boldsymbol{\lambda}^T]^T$ and similarly define $\mathbf{z}^* = [(\mathbf{x}^*)^T, (\boldsymbol{\lambda}^*)^T]^T$. Next, we define the following quadratic Lyapunov function $V(\mathbf{z}) = (\mathbf{z} - \mathbf{z}^*)^T P(\mathbf{z} - \mathbf{z}^*)$ with $P \succ 0$ defined by,

$$P = \begin{bmatrix} \eta c I & \eta A^T \\ \eta A & cI \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)} \tag{14}$$

where $c = 20\ell[\max(\frac{\rho\kappa_2}{\mu}, \frac{\ell}{\mu})]^2[\max(\frac{\eta}{\ell\rho}, \frac{\ell}{\mu})]^2\frac{\kappa_2}{\kappa_1}$.[4] Then, the results of Theorem 2 directly follows from the following property of $V(\mathbf{z})$,

$$\frac{d}{dt}V(\mathbf{z}) \le -\tau V(\mathbf{z}) \tag{15}$$

where $\tau = \frac{\eta\kappa_1}{2c} = \frac{\eta\kappa_1^2}{40\ell\kappa_2[\max(\frac{\rho\kappa_2}{\mu}, \frac{\ell}{\mu})]^2[\max(\frac{\eta}{\ell\rho}, \frac{\ell}{\mu})]^2}$. The rest of the section will be devoted to proving (15). To prove (15), we write the Aug-PDGD (9) in a "linear" form. In addition to Lemma 1 we need the following Lemma.

Lemma 3. *For any j and $\mathbf{z} = [\mathbf{x}^T, \boldsymbol{\lambda}^T]^T \in \mathbb{R}^{n+m}$, we have there exists $\gamma_j(\mathbf{z}) \in [0, 1]$ that depends on $\mathbf{z}$ s.t.*

$$\begin{aligned} &\nabla_{\mathbf{x}} H_\rho(\mathbf{a}_j^T\mathbf{x} - b_j, \lambda_j) - \nabla_{\mathbf{x}} H_\rho(\mathbf{a}_j^T\mathbf{x}^* - b_j, \lambda_j^*) \\ &= \gamma_j(\mathbf{z})\rho\mathbf{a}_j^T(\mathbf{x} - \mathbf{x}^*)\mathbf{a}_j + \gamma_j(\mathbf{z})(\lambda_j - \lambda_j^*)\mathbf{a}_j \\ &\nabla_{\boldsymbol{\lambda}} H_\rho(\mathbf{a}_j^T\mathbf{x} - b_j, \lambda_j) - \nabla_{\boldsymbol{\lambda}} H_\rho(\mathbf{a}_j^T\mathbf{x}^* - b_j, \lambda_j^*) \\ &= \gamma_j(\mathbf{z})\mathbf{a}_j^T(\mathbf{x} - \mathbf{x}^*)\mathbf{e}_j + \frac{1}{\rho}(\gamma_j(\mathbf{z}) - 1)(\lambda_j - \lambda_j^*)\mathbf{e}_j \end{aligned}$$

Proof. The lemma directly follows from that for any $y, y^* \in \mathbb{R}$, there exists some $\gamma \in [0, 1]$, depending on y, y^* s.t. $\max(y, 0) - \max(y^*, 0) = \gamma(y - y^*)$. To see this, when $y \ne y^*$, set $\gamma = \frac{\max(y,0) - \max(y^*,0)}{y - y^*}$; otherwise, set $\gamma = 0$. □

[3]We have $P \succ 0$ as long as $c^2 > \eta\kappa_2$, which is satisfied by our choice of c.

[4]We have $P \succ 0$ as long as $c^2 > \eta\kappa_2$, which is satisfied by our choice of c.

For any $\mathbf{z} \in \mathbb{R}^{n+m}$, we define notation $\Gamma(\mathbf{z}) = \text{diag}(\gamma_1(\mathbf{z}), \ldots, \gamma_m(\mathbf{z})) \in \mathbb{R}^{m\times m}$, where $\gamma_j(\mathbf{z})$ is from Lemma 3. With notation $\Gamma(\mathbf{z})$, we can then rewrite the Aug-PDGD (9a) as

$$\begin{aligned}\dot{\mathbf{x}} &= -(\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) - \nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)) \\ &= -(\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)) \\ &\quad - \sum_{j=1}^{m} \left[\nabla_{\mathbf{x}} H_\rho(\mathbf{a}_j^T \mathbf{x} - b_j, \lambda_j) - \nabla_{\mathbf{x}} H_\rho(\mathbf{a}_j^T \mathbf{x}^* - b_j, \lambda_j^*) \right] \\ &= -B(\mathbf{x})(\mathbf{x} - \mathbf{x}^*) - \rho A^T \Gamma(\mathbf{z}) A(\mathbf{x} - \mathbf{x}^*) - A^T \Gamma(\mathbf{z})(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)\end{aligned}$$

where $\mu I \preceq B(\mathbf{x}) \preceq \ell I$ (Lemma 1). We then rewrite (9b),

$$\begin{aligned}\dot{\boldsymbol{\lambda}} &= \eta \nabla_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) - \eta \nabla_{\boldsymbol{\lambda}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \\ &= \eta \sum_{j=1}^{m} \left[\nabla_{\boldsymbol{\lambda}} H_\rho(\mathbf{a}_j^T \mathbf{x} - b_j, \lambda_j) - \nabla_{\boldsymbol{\lambda}} H_\rho(a_j^T \mathbf{x}^* - b_j, \lambda_j^*) \right] \\ &= \eta \Gamma(\mathbf{z}) A(\mathbf{x} - \mathbf{x}^*) + \frac{\eta}{\rho}(\Gamma(\mathbf{z}) - I)(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)\end{aligned}$$

Then, the Aug-PDGD (9) can be written as,

$$\begin{aligned}\dot{\mathbf{z}} &= \begin{bmatrix} -B(\mathbf{x}) - \rho A^T \Gamma(\mathbf{z}) A & -A^T \Gamma(\mathbf{z}) \\ \eta \Gamma(\mathbf{z}) A & \frac{\eta}{\rho}(\Gamma(\mathbf{z}) - I) \end{bmatrix} (\mathbf{z} - \mathbf{z}^*) \\ &:= G(\mathbf{z})(\mathbf{z} - \mathbf{z}^*). \end{aligned} \tag{16}$$

Then, $\frac{d}{dt} V(\mathbf{z})$ can be written as

$$\begin{aligned}\frac{d}{dt} V(\mathbf{z}) &= \dot{\mathbf{z}}^T P(\mathbf{z} - \mathbf{z}^*) + (\mathbf{z} - \mathbf{z}^*)^T P \dot{\mathbf{z}} \\ &= (\mathbf{z} - \mathbf{z}^*)^T (G(\mathbf{z})^T P + P G(\mathbf{z}))(\mathbf{z} - \mathbf{z}^*) \end{aligned} \tag{17}$$

Therefore, to prove (15), it is sufficient to show the following Lemma, whose proof is in Appendix-B.

Lemma 4. *For any $\mathbf{z} \in \mathbb{R}^{n+m}$, we have*

$$G(\mathbf{z})^T P + P G(\mathbf{z}) \preceq -\tau P$$

Lemma 4 and (17) lead to (15), concluding the proof.

C. Discrete Time Primal-Dual Gradient Algorithm

Lastly, we briefly discuss the stability of the discretization of (Aug-)PDGD. It is known that the Euler discretization of an exponentially stable dynamical system possesses geometric convergence speed [23], [24], provided the discretization step size is small enough. For completeness, we provide the following Lemma 5, whose proof can be found in our online report [31, Appendix-F].

Lemma 5. *Consider a continuous-time dynamical system $\dot{\mathbf{z}} = F(\mathbf{z})$ where F is ν-Lipschitz continuous. Suppose $\mathbf{z}^*$ is an equilibrium point and there exists positive definite matrix P, constant $\tau > 0$, and Lyapunov function $V(\tilde{\mathbf{z}}) = \tilde{\mathbf{z}}^T P \tilde{\mathbf{z}}$ such that $\frac{d}{dt} V(\tilde{\mathbf{z}}) \leq -\tau V(\tilde{\mathbf{z}})$, where $\tilde{\mathbf{z}} := \mathbf{z} - \mathbf{z}^*$. Then its Euler discretization with step size $\delta > 0$,*

$$\mathbf{y}(k+1) = \mathbf{y}(k) + \delta F(\mathbf{y}(k))$$

satisfies $\|\mathbf{y}(k) - \mathbf{z}^*\| \leq C(e^{-\frac{\tau\delta}{2}} + \frac{\kappa_P \nu^2 \delta^2}{2})^k$, where κ_P is the condition number of matrix P, and $C > 0$ is some constant that depends on ν, τ, P and $\|\mathbf{y}(0) - \mathbf{z}^*\|$. Further, $e^{-\frac{\tau\delta}{2}} + \frac{\kappa_P \nu^2 \delta^2}{2} < 1$ for small enough δ.

Based on the proof in Section III-A and III-B, both PDGD (6) and Aug-PDGD (9) satisfy the conditions in Lemma 5. Hence the discretized versions converge geometrically fast for small enough discretization step size δ.

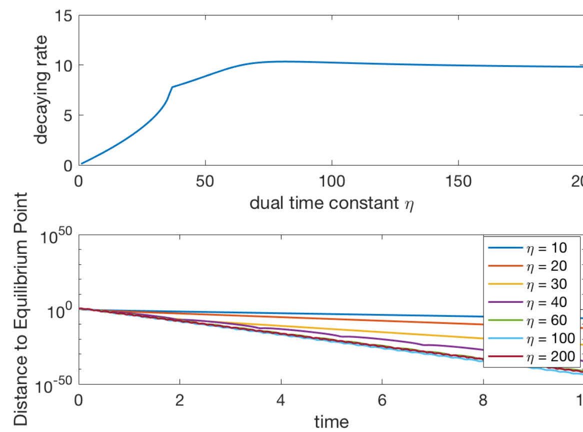

Fig. 1: Simulation of PDGD for the equality constrained case. Upper: decaying rate as a function of η. Lower: distance to the equilibrium point for some values of η.

IV. Illustrative Examples

A. Equality Constrained Case

We numerically study PDGD with affine equality constraints and quadratic cost functions. We let $n = 5$, $m = 2$, $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T W \mathbf{x}$, where $W = 10I + W_0 W_0^T$, and W_0 is a n-by-n Gaussian random matrix. A and $\mathbf{b}$ are also Gaussian random matrices (or vectors). Since the cost is quadratic, the PDGD (6) becomes an Linear Time-Invariant (LTI) system and we can determine the PDGD decaying rate by numerically calculating the eigenvalues of the resulting LTI system. We plot the decaying rate as a function of η in the upper plot of Fig. 1. We also simulate the PDGD for a selected number of η's, and plot the distance to equilibrium point as a function of time in the lower plot of Fig. 1. In both plots, we observe that increasing η beyond a certain threshold does not lead to faster decaying rate, an interesting phenomenon that may be worth further studying.

B. Inequality Constrained Case

We numerically run the Aug-PDGD on a problem of size $n = 50, m = 40$. We use the loss for logistic regression [32] (with synthetic data) as our cost function f. For the affine inequality constraint $A\mathbf{x} \leq \mathbf{b}$, every entry of A, $\mathbf{b}$ is generated independently from standard normal distribution. We fix $\rho = 1$ but try different η's, and show the results in Fig. 2. Similar to the equality constrained case, here we observe that when η is large, the decaying rate doesn't increase with η.

V. Conclusions

In this paper, we study the primal-dual gradient dynamics for optimization with strongly convex and smooth objective and affine equality or inequality constraints. We prove the global exponential stability of PDGD. We also give explicit bounds on the decaying rates. Future work include 1) providing tighter bounds on the decaying rates, especially for

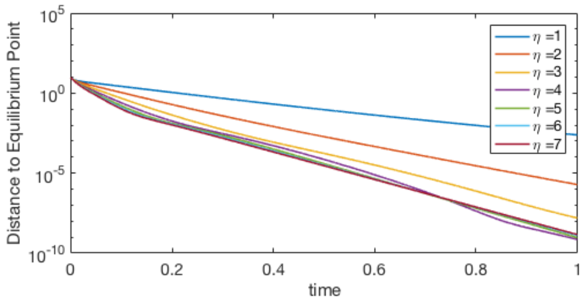

Fig. 2: Simulation of Aug-PDGD for the inequality constrained case under different values of η.

the inequality constrained case; 2) relaxing Assumption 2 for the inequality constrained case.

References

[1] T. Kose, "Solutions of saddle value problems by differential equations," *Econometrica, Journal of the Econometric Society*, pp. 59–70, 1956.

[2] K. J. Arrow, L. Hurwicz, H. Uzawa, and H. B. Chenery, *Studies in linear and non-linear programming*. Stanford University Press, 1958.

[3] C. Zhao, U. Topcu, N. Li, and S. Low, "Design and stability of load-side primary frequency control in power systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1177–1189, 2014.

[4] N. Li, C. Zhao, and L. Chen, "Connecting automatic generation control and economic dispatch from an optimization view," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 3, pp. 254–264, 2016.

[5] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[6] J. Chen and V. K. Lau, "Convergence analysis of saddle point problems in time varying wireless systemscontrol theoretical approach," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 443–452, 2012.

[7] J. Wang and N. Elia, "A control perspective for centralized and distributed convex optimization," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on.* IEEE, 2011, pp. 3800–3805.

[8] S. K. Niederländer and J. Cortés, "Distributed coordination for nonsmooth convex optimization via saddle-point dynamics," *arXiv preprint arXiv:1606.09298*, 2016.

[9] B. Gharesifard and J. Cortés, "Distributed convergence to nash equilibria in two-network zero-sum games," *Automatica*, vol. 49, no. 6, pp. 1683–1692, 2013.

[10] H. Uzawa, "Iterative methods in concave programming," *Studies in linear and non-linear programming*, 1958.

[11] B. Polyak, "Iterative methods using lagrange multipliers for solving extremal problems with constraints of the equation type," *USSR Computational Mathematics and Mathematical Physics*, vol. 10, no. 5, pp. 42–52, 1970.

[12] E. Golshtein, "Generalized gradient method for finding saddlepoints," *Matekon*, vol. 10, no. 3, pp. 36–52, 1974.

[13] I. Zabotin, "A subgradient method for finding a saddle point of a convex-concave function," *Issled. Prikl. Mat*, vol. 15, pp. 6–12, 1988.

[14] M. Kallio and C. H. Rosa, "Large-scale convex optimization via saddle point computation," *Operations Research*, vol. 47, no. 1, pp. 93–101, 1999.

[15] M. Kallio and A. Ruszczynski, "Perturbation methods for saddle point computation," 1994.

[16] G. Korpelevich, "The extragradient method for finding saddle points and other problems," *Matecon*, vol. 12, pp. 747–756, 1976.

[17] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of optimization theory and applications*, vol. 142, no. 1, pp. 205–228, 2009.

[18] 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.

[19] T. Holding and I. Lestas, "On the convergence to saddle points of concave-convex functions, the gradient method and emergence of oscillations," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on.* IEEE, 2014, pp. 1143–1148.

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

[21] A. Cherukuri, E. Mallada, S. Low, and J. Cortés, "The role of convexity on saddle-point dynamics: Lyapunov function and robustness," *IEEE Transactions on Automatic Control*, 2017.

[22] A. Cherukuri, B. Gharesifard, and J. Cortes, "Saddle-point dynamics: conditions for asymptotic stability of saddle points," *SIAM Journal on Control and Optimization*, vol. 55, no. 1, pp. 486–511, 2017.

[23] A. M. Stuart, "Numerical analysis of dynamical systems," *Acta numerica*, vol. 3, pp. 467–572, 1994.

[24] H. J. Stetter, *Analysis of discretization methods for ordinary differential equations*. Springer, 1973, vol. 23.

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

[26] M. Benzi, G. H. Golub, and J. Liesen, "Numerical solution of saddle point problems," *Acta numerica*, vol. 14, pp. 1–137, 2005.

[27] S.-Q. Shen, T.-Z. Huang, and J. Yu, "Eigenvalue estimates for preconditioned nonsymmetric saddle point matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 5, pp. 2453–2476, 2010.

[28] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. Siam, 1970, vol. 30.

[29] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[30] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented lagrangian method for nonsmooth composite optimization," *arXiv preprint arXiv:1610.04514*, 2016.

[31] G. Qu and N. Li, "On the exponential stability of primal-dual gradient dynamics," *arXiv preprint arXiv:1803.01825*, 2018.

[32] (2012) Logistic regression. [Online]. Available: http://www.stat.cmu.edu/ cshalizi/uADA/12/lectures/ch12.pdf

Appendix

A. Proof of Lemma 2

Recall the definition of $G(\mathbf{z})$ in (12),

$$G(\mathbf{z}) = \begin{bmatrix} -B(\mathbf{x}) & -A^T \\ \eta A & 0 \end{bmatrix}$$

It can be seen that $G(\mathbf{z})$ depends on $\mathbf{z}$ through $B(\mathbf{x})$, and $B(\mathbf{x})$ satisfies $\mu I \preceq B(\mathbf{x}) \preceq \ell I$. In the remaining of this section, we will drop the dependence of $G(\mathbf{z})$ and $B(\mathbf{x})$ on $\mathbf{z}$ and $\mathbf{x}$, and prove $G^T P + PG \preceq -\tau P$, for *any* symmetric B satisfying $\mu I \preceq B \preceq \ell I$. Let $Q = -G^T P - PG$, then,

$$\begin{aligned} Q - \tau P &= \begin{bmatrix} 2\eta c B - 2\eta^2 A^T A - \eta^2\kappa_1 I & \eta B A^T - \frac{\eta^2\kappa_1}{c} A^T \\ \eta A B - \frac{\eta^2\kappa_1}{c} A & 2\eta A A^T - \eta\kappa_1 I \end{bmatrix} \\ &\succeq \begin{bmatrix} 2\eta c B - 2\eta^2 \kappa_2 I - \eta^2\kappa_1 I & \eta B A^T - \frac{\eta^2\kappa_1}{c} A^T \\ \eta A B - \frac{\eta^2\kappa_1}{c} A & \eta A A^T \end{bmatrix} \end{aligned}$$

where we have used $AA^T \succeq \kappa_1 I$, $A^T A \preceq \kappa_2 I$. We will next use the Schur complement argument. Consider

$$\begin{aligned} &(\eta B A^T - \frac{\eta^2\kappa_1}{c} A^T)(\eta A A^T)^{-1}(\eta A B - \frac{\eta^2\kappa_1}{c} A) \\ &= \eta B A^T (AA^T)^{-1} A B - \frac{\eta^2\kappa_1}{c}(B A^T (AA^T)^{-1} A \\ &\quad + A^T (AA^T)^{-1} A B) + \frac{\eta^3\kappa_1^2}{c^2} A^T (AA^T)^{-1} A \\ &\preceq \eta \ell B + \frac{\eta^2\kappa_1}{c} 2\ell I + \frac{\eta^3\kappa_1^2}{c^2} I \end{aligned}$$

where we have used $A^T(AA^T)^{-1}A \preceq I$. Recall that $c = 4\max(\ell, \frac{\eta\kappa_2}{\mu})$, $\kappa_1 \le \kappa_2$ and $\mu \le l$. Then we have, i) $\frac{3}{4}\eta c B \succeq \frac{3}{4}\eta c\mu I \succeq 3\eta^2\kappa_2 I \succeq 2\eta^2\kappa_2 I + \eta^2\kappa_1 I$, ii) $\frac{1}{4}\eta c B \succeq \eta\ell B$, iii) $\frac{1}{2}\eta c B \succeq \frac{1}{2}\eta c\mu I \succeq \frac{\eta^2}{c} 2\ell\kappa_1 I$, and iv)$\frac{1}{2}\eta c B \succeq \frac{1}{2}\eta c\mu I \succeq \frac{\eta^3\kappa_1^2}{c^2} I$. Summing them up, we have

$$2\eta c B \succeq 2\eta^2\kappa_2 I + \eta^2\kappa_1 I + \eta\ell B + \frac{\eta^2}{c} 2\ell\kappa_1 I + \frac{\eta^3\kappa_1^2}{c^2} I$$

$$\succeq 2\eta^2\kappa_2 I + \eta^2\kappa_1 I + (\eta BA^T - \frac{\eta^2\kappa_1}{c}A^T)(\eta AA^T)^{-1}(\eta AB - \frac{\eta^2\kappa_1}{c}A)$$

As a result, we have $Q - \tau P \succeq 0$.

B. Proof of Lemma 4

Recall the definition of $G(\mathbf{z})$ in (16),

$$G(\mathbf{z}) = \begin{bmatrix} -B(\mathbf{x}) - \rho A^T\Gamma(\mathbf{z})A & -A^T\Gamma(\mathbf{z}) \\ \eta\Gamma(\mathbf{z})A & \frac{\eta}{\rho}(\Gamma(\mathbf{z}) - I) \end{bmatrix}.$$

It can be seen that $G(\mathbf{z})$ depends on the state $\mathbf{z}$ through $B(\mathbf{x})$, $\Gamma(\mathbf{z})$. Note for any $\mathbf{z} \in \mathbb{R}^{n+m}$, $B(\mathbf{x})$ is a symmetric matrix satisfying $\mu I \preceq B(\mathbf{x}) \preceq \ell I$, and $\Gamma(\mathbf{z})$ is a diagonal matrix with each entry in $[0,1]$. In the following, to simplify notation, we will drop the dependence of G, B, Γ on $\mathbf{z}$, and prove $G^T P + PG \preceq -\tau P$ for any symmetric B satisfing $\mu I \preceq B \preceq \ell I$ and for any diagonal Γ with each entry bounded in $[0,1]$. Let $Q = -G^T P - PG$, and

$$\tilde{Q} = Q - \tau P = \begin{bmatrix} Q_1 & Q_3 \\ Q_3^T & Q_2 \end{bmatrix}$$

After straightforward calculations, we have

$$Q_1 = 2\eta cB + 2(\eta c\rho - \eta^2)A^T\Gamma A - \frac{\eta^2\kappa_1}{2}I$$

$$Q_2 = \eta(\Gamma AA^T + AA^T\Gamma) + \frac{2\eta c}{\rho}(I - \Gamma) - \frac{\eta\kappa_1}{2}I$$

$$\begin{aligned} Q_3 &= \eta BA^T + \eta\rho A^T\Gamma AA^T + \frac{\eta^2}{\rho}A^T(I-\Gamma) - \frac{\eta^2\kappa_1}{2c}A^T \\ &= \eta(\underbrace{B + \overbrace{\rho A^T\Gamma A - \frac{\eta\kappa_1}{2c}I}^{Q_5}}_{Q_4})A^T + \eta\cdot\underbrace{\frac{\eta}{\rho}A^T(I-\Gamma)}_{Q_6} \end{aligned}$$

Using the Schur complement argument, to prove $\tilde{Q} \succeq 0$ it suffices to prove $Q_2 \succ 0$, $Q_1 - Q_3 Q_2^{-1}Q_3^T \succeq 0$. To show this, we will first lower bound Q_2, then upper bound $Q_3Q_2^{-1}Q_3^T$, next lower bound Q_1 and finally show $Q_1 - Q_3Q_2^{-1}Q_3^T \succeq 0$.

Lower bounding Q_2. We will use the following lemma, whose proof is deferred to Appendix-C.

Lemma 6. *If $c \geq \kappa_2\rho$, as long as Γ is a diagonal matrix with each entry bounded in $[0,1]$, we have $\eta(\Gamma AA^T + AA^T\Gamma) + \frac{2\eta c}{\rho}(I - \Gamma) \succeq \frac{3}{2}\eta AA^T$.*

Using Lemma 6, we have $Q_2 \succeq \frac{3}{2}\eta AA^T - \frac{\eta\kappa_1}{2}I \succeq \eta AA^T$.

Upper bounding $Q_3Q_2^{-1}Q_3^T$. Using the lower bound on Q_2, we calculate

$$\begin{aligned} Q_3Q_2^{-1}Q_3^T &\preceq \frac{1}{\eta}Q_3(AA^T)^{-1}Q_3^T \\ &= \eta Q_4A^T(AA^T)^{-1}AQ_4^T + \eta Q_4A^T(AA^T)^{-1}Q_6^T \\ &\quad + \eta Q_6(AA^T)^{-1}AQ_4^T + \eta Q_6(AA^T)^{-1}Q_6^T \\ &\preceq \eta Q_4Q_4^T + 2\eta\|Q_4A^T(AA^T)^{-1}Q_6^T\|I \\ &\quad + \eta\|Q_6(AA^T)^{-1}Q_6^T\|I \end{aligned} \tag{18}$$

where in the last inequality we have used $A^T(AA^T)^{-1}A \preceq I$. To further bound $Q_3Q_2^{-1}Q_3^T$, we use the following,

$$\begin{aligned} 2\eta\|Q_4A^T(AA^T)^{-1}Q_6^T\| &\leq 2\eta\|Q_4\|\|A^T\|\|(AA^T)^{-1}\|\|Q_6\| \\ &\leq 2\frac{\eta^2}{\rho}(\ell + \rho\kappa_2 + \frac{\eta\kappa_1}{2c})\frac{\kappa_2}{\kappa_1} := h_1(c) \end{aligned} \tag{19}$$

where we intentionally write the last quantity as a function $h_1(c)$ depending on c for reasons to be clear later. We further bound

$$\eta\|Q_6(AA^T)^{-1}Q_6^T\| \leq \eta\|Q_6\|^2\|(AA^T)^{-1}\| \leq \frac{\eta^3}{\rho^2}\frac{\kappa_2}{\kappa_1} := h_2(c) \tag{20}$$

Then, we bound $\eta Q_4Q_4^T$ as

$$\begin{aligned} \eta Q_4Q_4^T &= \eta(B + Q_5)^2 \preceq \eta B^2 + 2\eta\|B\|\|Q_5\|I + \eta\|Q_5\|^2 I \\ &\preceq \eta\ell B + \underbrace{(2\eta\ell(\rho\kappa_2 + \frac{\eta\kappa_1}{2c}) + \eta(\rho\kappa_2 + \frac{\eta\kappa_1}{2c})^2)}_{:=h_3(c)}I \end{aligned} \tag{21}$$

Combining (18) (19) (20) (21), we have $Q_3Q_2^{-1}Q_3^T \preceq \eta\ell B + (h_1(c) + h_2(c) + h_3(c))I$.

Lower bounding Q_1. It is easy to obtain,

$$Q_1 \succeq 2\eta cB - \underbrace{(2\eta^2\kappa_2 + \frac{\eta^2\kappa_1}{2})}_{:=h_4(c)}I \tag{22}$$

Proving $Q_1 - Q_3Q_2^{-1}Q_3^T \succeq 0$. Combining the above bounds on Q_1 and $Q_3Q_2^{-1}Q_3^T$, we have,

$$\begin{aligned} &Q_1 - Q_3Q_2^{-1}Q_3^T \\ &\succeq (2\eta c - \eta\ell)B - (h_1(c) + h_2(c) + h_3(c) + h_4(c))I \\ &\succeq \left[2\eta\mu c - (\eta\ell\mu + h_1(c) + h_2(c) + h_3(c) + h_4(c))\right]I \end{aligned}$$

Since $h_1(c) + h_2(c) + h_3(c) + h_4(c)$ is a strictly decreasing function in c, we have $Q_1 - Q_3Q_2^{-1}Q_3^T \succeq 0$ as long as c is sufficiently large. We can verify that our selection of c is large enough s.t. $Q_1 - Q_3Q_2^{-1}Q_3^T \succeq 0$. Due to space limit, we omit the details. Therefore, $Q \succeq \tau P$.

C. Proof of Lemma 6

Recall that $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_m)$ with each $\gamma_i \in [0,1]$. The matrix of interest is $M(\gamma_1, \ldots, \gamma_m) = \eta(\Gamma AA^T + AA^T\Gamma) + \frac{2\eta c}{\rho}(I - \Gamma)$. It is easy to check $M(\gamma_1, \ldots, \gamma_m)$ is a convex combination of 2^m matrices $\{M(b_1, \ldots, b_m) : b_i = 0 \text{ or } 1\}$. Therefore, to prove the lower bound for $M(\gamma_1, \ldots, \gamma_m)$, without loss of generality, we only have to prove that for $k = 0, \ldots, m$

$$M_k = M(\underbrace{1, 1, \ldots, 1}_{k \text{ entries}}, 0, \ldots, 0) \succeq \frac{3}{2}\eta AA^T \tag{23}$$

Notice that $M_0 = \frac{2\eta c}{\rho}I \succeq \frac{3}{2}\eta AA^T$, $M_m = 2\eta AA^T$, so (23) is true for $k = 0$ and m. Now assume $0 < k < m$. We write AA^T in block matrix form

$$AA^T = \begin{bmatrix} \Lambda_1 & \Lambda_3 \\ \Lambda_3^T & \Lambda_2 \end{bmatrix} \tag{24}$$

with $\Lambda_1 \in \mathbb{R}^{k\times k}, \Lambda_2 \in \mathbb{R}^{(m-k)\times(m-k)}, \Lambda_3 \in \mathbb{R}^{(k)\times(m-k)}$. Then, we can write M_k as

$$M_k = \begin{bmatrix} 2\eta\Lambda_1 & \eta\Lambda_3 \\ \eta\Lambda_3^T & 2\frac{\eta c}{\rho}I \end{bmatrix} \succeq \begin{bmatrix} 2\eta\Lambda_1 & \eta\Lambda_3 \\ \eta\Lambda_3^T & 2\eta\Lambda_2 \end{bmatrix} \succeq \frac{3}{2}\eta AA^T$$

where we have used the fact that $\Lambda_2 \preceq \|\Lambda_2\|I \preceq \kappa_2 I \preceq \frac{c}{\rho}I$ (using $c \geq \rho\kappa_2$).