A Primal-Dual Type Algorithm with the $O(1/t)$ Convergence Rate for Large Scale Constrained Convex Programs

Hao Yu and Michael J. Neely

***Abstract*— This paper considers large scale constrained convex programs. These are often difficult to solve by interior point methods or other Newton-type methods due to the prohibitive computation and storage complexity for Hessians or matrix inversions. Instead, large scale constrained convex programs are often solved by gradient based methods or decomposition based methods. The conventional primal-dual subgradient method, also known as the Arrow-Hurwicz-Uzawa subgradient method, is a low complexity algorithm with the $O(1/\sqrt{t})$ convergence rate, where t is the number of iterations. If the objective and constraint functions are separable, the Lagrangian dual type method can decompose a large scale convex program into multiple parallel small scale convex programs. The classical dual gradient algorithm is an example of Lagrangian dual type methods and has convergence rate $O(1/\sqrt{t})$. Recently, the authors of the current paper proposed a new Lagrangian dual type algorithm with faster $O(1/t)$ convergence. However, if the objective or constraint functions are not separable, each iteration requires to solve a large scale unconstrained convex program, which can have huge complexity. This paper proposes a new primal-dual type algorithm, which only involves simple gradient updates at each iteration and has $O(1/t)$ convergence.**

I. Introduction

Fix positive integers n and m, which are typically large. Consider the general constrained convex program:

$$\begin{aligned} \text{minimize:} \quad & f(\mathbf{x}) && (1)\\ \text{such that:} \quad & g_k(\mathbf{x}) \leq 0, \forall k \in \{1,2,\ldots,m\} && (2)\\ & \mathbf{x} \in \mathcal{X} && (3)\end{aligned}$$

where set $\mathcal{X} \subseteq \mathbb{R}^n$ is a compact convex set; function $f(\mathbf{x})$ is convex and smooth on $\mathcal{X}$; and functions $g_k(\mathbf{x}), \forall k \in \{1,2,\ldots,m\}$ are convex, smooth and Lipschitz continuous on $\mathcal{X}$. Denote the stacked vector of multiple functions $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})$ as $\mathbf{g}(\mathbf{x}) = \big[g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})\big]^T$. The Lipschitz continuity of each $g_k(\mathbf{x})$ implies that $\mathbf{g}(\mathbf{x})$ is Lipschitz continuous on $\mathcal{X}$. Throughout this paper, we use $\|\cdot\|$ to represent the Euclidean norm and have the following assumptions on convex program (1)-(3):

Assumption 1 (Basic Assumptions):

- There exists a (possibly non-unique) optimal solution $\mathbf{x}^* \in \mathcal{X}$ that solves convex program (1)-(3).
- There exists $L_f \geq 0$ such that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_f\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, i.e., $f(\mathbf{x})$ is smooth with modulus L_f. For each $k \in \{1,2,\ldots,m\}$, there exists $L_{g_k} \geq 0$ such that $\|\nabla g_k(\mathbf{x}) - \nabla g_k(\mathbf{y})\| \leq L_{g_k}\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, i.e., $g_k(\mathbf{x})$ is smooth with modulus L_{g_k}. Denote $\mathbf{L_g} = [L_{g_1}, \ldots, L_{g_m}]^T$.
- There exists $\beta \geq 0$ such that $\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| \leq \beta\|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$, i.e., $\mathbf{g}(\mathbf{x})$ is Lipschitz continuous with modulus β.
- There exists $C \geq 0$ such that $\|\mathbf{g}(\mathbf{x})\| \leq C, \forall \mathbf{x} \in \mathcal{X}$.
- There exists $R \geq 0$ such that $\|\mathbf{x}-\mathbf{y}\| \leq R, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$.

Note that the existence of C follows from the continuity of $\mathbf{g}(\mathbf{x})$ and the compactness of set $\mathcal{X}$. The existence of R follows from the compactness of set $\mathcal{X}$.

Assumption 2 (Existence of Lagrange multipliers): There exists a Lagrange multiplier vector $\boldsymbol{\lambda}^* = [\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*] \geq \mathbf{0}$ attaining the strong duality for problem (1)-(3), i.e.,

$$q(\boldsymbol{\lambda}^*) = \min_{\mathbf{x}\in\mathcal{X}} \{f(\mathbf{x}) : g_k(\mathbf{x}) \leq 0, \forall k \in \{1,2,\ldots,m\}\},$$

where $q(\boldsymbol{\lambda}) = \min_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x}) + \sum_{k=1}^m \lambda_k g_k(\mathbf{x})\}$ is the *Lagrangian dual function* of problem (1)-(3).

Assumption 2 is a mild condition. For example, it is implied by the *Slater condition* for convex programs [1].

A. Large Scale Convex Programs

In general, convex program (1)-(3) can be solved via interior point methods (or other Newton type methods) which involve the computation of Hessians and matrix inversions at each iteration. The associated computation complexity and memory space complexity at each iteration is between $O(n^2)$ and $O(n^3)$, which is prohibitive when n is extremely large. For example, if $n = 10^5$ and each floating point number uses 4 bytes, then 40 Gbytes of memory is required even to save the Hessian at each iteration. Thus, large scale convex programs are usually solved by gradient based methods or decomposition based methods.

B. The Primal-Dual Subgradient Method

The primal-dual subgradient method, also known as the Arrow-Hurwicz-Uzawa Subgradient Method, applied to convex program (1)-(3) is described in Algorithm 1. The updates of $\mathbf{x}(t)$ and $\boldsymbol{\lambda}(t)$ only involve the computation of gradient and simple projection operations, which are much simpler than the computation of Hessians and matrix inversions for extremely large n. Thus, compared with the interior point methods, the primal-dual subgradient algorithm has lower complexity computations at each iteration and hence is more suitable to large scale convex programs. However, the

The authors are with the Electrical Engineering department at the University of Southern California, Los Angeles, CA.

convergence rate[1] of Algorithm 1 is only $O(1/\sqrt{t})$, where t is the number of iterations [2].

Algorithm 1 The Primal-Dual Subgradient Algorithm

Let $c > 0$ be a constant step size. Choose any $\mathbf{x}(0) \in \mathcal{X}$. Initialize Lagrangian multipliers $\lambda_k(0) = 0, \forall k \in \{1, 2, \ldots, m\}$. At each iteration $t \in \{1, 2, \ldots\}$, observe $\mathbf{x}(t-1)$ and $\boldsymbol{\lambda}(t-1)$ and do the following:

- Choose $\mathbf{x}(t) = \mathcal{P}_{\mathcal{X}}\big[\mathbf{x}(t-1) - c\nabla f(\mathbf{x}(t-1)) - c\sum_{k=1}^{m} \lambda_k(t-1)\nabla g_k(\mathbf{x}(t-1))\big]$, where $\mathcal{P}_{\mathcal{X}}[\cdot]$ is the projection onto convex set $\mathcal{X}$.
- Update Lagrangian multipliers $\lambda_k(t) = [\lambda_k(t-1) + cg_k(\mathbf{x}(t-1))]_0^{\lambda_k^{\max}}, \forall k \in \{1, 2, \ldots, m\}$, where $\lambda_k^{\max} > \lambda_k^*$ and $[\cdot]_0^{\max}$ is the projection onto interval $[0, \lambda_k^{\max}]$.
- Update the running averages $\overline{\mathbf{x}}(t+1) = \frac{1}{t}\sum_{\tau=0}^{t} \mathbf{x}(\tau) = \overline{\mathbf{x}}(t)\frac{t}{t+1} + \mathbf{x}(t)\frac{1}{t+1}$.

C. Lagrangian Dual Type Methods

The classical dual subgradient algorithm is a Lagrangian dual type iterative method that approaches optimality for strictly convex programs [3]. A modification of the classical dual subgradient algorithm that averages the resulting sequence of primal estimates can solve general convex programs and has an $O(1/\sqrt{t})$ convergence rate [4], [5], [6]. The dual subgradient algorithm with primal averaging is suitable to large scale convex programs because the updates of each component $x_i(t)$ are independent and parallel if functions $f(\mathbf{x})$ and $g_k(\mathbf{x})$ in convex program (1)-(3) are separable with respect to each component (or block) of $\mathbf{x}$, e.g., $f(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i)$ and $g_k(\mathbf{x}) = \sum_{i=1}^{n} g_{k,i}(x_i)$.

Recently, a new Lagrangian dual type algorithm with convergence rate $O(1/t)$ for general convex programs is proposed in [7]. This algorithm can solve convex program (1)-(3) following the steps described in Algorithm 2.

Similar to the dual subgradient algorithm with primal averaging, Algorithm 2 can decompose the updates of $\mathbf{x}(t)$ into smaller independent subproblems if functions $f(\mathbf{x})$ and $g_k(\mathbf{x})$ are separable. Moreover, Algorithm 2 has $O(1/t)$ convergence, which is faster than the primal-dual subgradient or the dual subgradient algorithm with primal averaging.

However, if $f(\mathbf{x})$ or $g_k(\mathbf{x})$ are not separable, each update of $\mathbf{x}(t)$ requires to solve a set constrained convex program. If the dimension n is large, such a set constrained convex program should be solved via a gradient based method instead of a Newton method. However, the gradient based method for set constrained convex programs is an iterative technique and involves at least one projection operation at each iteration.

[1]In this paper, we say that the primal dual subgradient algorithm and the dual subgradient algorithm have an $O(1/\sqrt{t})$ convergence rate in the sense that they achieve an ϵ-approximate solution with $O(1/\epsilon^2)$ iterations by using an $O(\epsilon)$ step size. The error of those algorithms does not necessarily continue to decay after the ϵ-approximate solution is reached. In contrast, the algorithm in the current paper has a faster $O(1/t)$ convergence and this holds for all time t, so that error goes to zero as the number of iterations increases.

Algorithm 2 Algorithm 1 in [7]

Let $\alpha > 0$ be a constant parameter. Choose any $\mathbf{x}(-1) \in \mathcal{X}$. Initialize virtual queues $Q_k(0) = \max\{0, -g_k(\mathbf{x}(-1))\}, \forall k \in \{1, 2, \ldots, m\}$. At each iteration $t \in \{0, 1, 2, \ldots\}$, observe $\mathbf{x}(t-1)$ and $\mathbf{Q}(t)$ and do the following:

- Choose $\mathbf{x}(t) = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \Big\{f(\mathbf{x}) + [\mathbf{Q}(t) + \mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{g}(\mathbf{x}) + \alpha\|\mathbf{x} - \mathbf{x}(t-1)\|^2\Big\}$.
- Update virtual queue vector $\mathbf{Q}(t)$ via $Q_k(t+1) = \max\{-g_k(\mathbf{x}(t)), Q_k(t) + g_k(\mathbf{x}(t))\}, \forall k \in \{1, 2, \ldots, m\}$.
- Update the running averages via $\overline{\mathbf{x}}(t+1) = \frac{1}{t+1}\sum_{\tau=0}^{t} \mathbf{x}(\tau) = \overline{\mathbf{x}}(t)\frac{t}{t+1} + \mathbf{x}(t)\frac{1}{t+1}$.

D. New Algorithm

Consider large scale convex programs with non-separable $f(\mathbf{x})$ or $g_k(\mathbf{x})$, e.g., $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$. In this case, Algorithm 1 has convergence rate $O(1/\sqrt{t})$ using low complexity iterations; while Algorithm 2 has convergence rate $O(1/t)$ using high complexity iterations.

This paper proposes a new algorithm described in Algorithm 3 which combines the advantages of Algorithm 1 and Algorithm 2. The new algorithm modifies Algorithm 2 by changing the update of $\mathbf{x}(t)$ from a minimization problem to a simple projection. Meanwhile, the $O(1/t)$ convergence rate of Algorithm 2 is preserved in the new algorithm.

Algorithm 3 New Algorithm

Let $\gamma > 0$ be a constant step size. Choose any $\mathbf{x}(-1) \in \mathcal{X}$. Initialize virtual queues $Q_k(0) = \max\{0, -g_k(\mathbf{x}(-1))\}, \forall k \in \{1, 2, \ldots, m\}$. At each iteration $t \in \{0, 1, 2, \ldots\}$, observe $\mathbf{x}(t-1)$ and $\mathbf{Q}(t)$ and do the following:

- Define $\mathbf{d}(t) = \nabla f(\mathbf{x}(t-1)) + \sum_{k=1}^{m}[Q_k(t) + g_k(\mathbf{x}(t-1))]\nabla g_k(\mathbf{x}(t-1))$, which is the gradient of function $\phi(\mathbf{x}) = f(\mathbf{x}) + [\mathbf{Q}(t) + \mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{g}(\mathbf{x})$ at point $\mathbf{x} = \mathbf{x}(t-1)$. Choose $\mathbf{x}(t) = \mathcal{P}_{\mathcal{X}}\left[\mathbf{x}(t-1) - \gamma\mathbf{d}(t)\right]$, where $\mathcal{P}_{\mathcal{X}}[\cdot]$ is the projection onto convex set $\mathcal{X}$.
- Update virtual queue vector $\mathbf{Q}(t)$ via $Q_k(t+1) = \max\{-g_k(\mathbf{x}(t)), Q_k(t) + g_k(\mathbf{x}(t))\}, \forall k \in \{1, 2, \ldots, m\}$.
- Update the running averages $\overline{\mathbf{x}}(t+1) = \frac{1}{t+1}\sum_{\tau=0}^{t} \mathbf{x}(\tau) = \overline{\mathbf{x}}(t)\frac{t}{t+1} + \mathbf{x}(t)\frac{1}{t+1}$.

II. Preliminaries and Basic Analysis

This section presents useful preliminaries on convex analysis and important facts of Algorithm 3.

A. Preliminaries

Definition 1 (Lipschitz Continuity): Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Function $h : \mathcal{X} \to \mathbb{R}^m$ is said to be Lipschitz continuous on $\mathcal{X}$ with modulus L if there exists $L > 0$ such that $\|h(\mathbf{y}) - h(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

Definition 2 (Smooth Functions): Let $\mathcal{X} \subseteq \mathbb{R}^n$ and function $h(\mathbf{x})$ be continuously differentiable on $\mathcal{X}$. Function $h(\mathbf{x})$ is said to be smooth on $\mathcal{X}$ with modulus L if $\nabla h(\mathbf{x})$ is Lipschitz continuous on $\mathcal{X}$ with modulus L.

Note that linear function $h(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ is smooth with modulus 0. If a function $h(\mathbf{x})$ is smooth with modulus L, then $ch(\mathbf{x})$ is smooth with modulus cL for any $c > 0$.

Lemma 1 (Descent Lemma, Proposition A.24 in [3]): If h is smooth on $\mathcal{X}$ with modulus L, then $h(\mathbf{y}) \leq h(\mathbf{x}) + \nabla h(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}||\mathbf{y}-\mathbf{x}||^2$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

Definition 3 (Strongly Convex Functions): Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Function h is said to be strongly convex on $\mathcal{X}$ with modulus α if there exists a constant $\alpha > 0$ such that $h(\mathbf{x}) - \frac{1}{2}\alpha\|\mathbf{x}\|^2$ is convex on $\mathcal{X}$.

If $h(\mathbf{x})$ is convex and $\alpha > 0$, then $h(\mathbf{x}) + \alpha\|\mathbf{x}-\mathbf{x}_0\|^2$ is strongly convex with modulus 2α for any constant $\mathbf{x}_0$.

Lemma 2: Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set. Let function h be strongly convex with modulus α and $\mathbf{x}^{opt}$ be a global minimum of h on $\mathcal{X}$. Then, $h(\mathbf{x}^{opt}) \leq h(\mathbf{x}) - \frac{\alpha}{2}\|\mathbf{x}^{opt} - \mathbf{x}\|^2, \forall \mathbf{x} \in \mathcal{X}$.

Proof: A special case when h is differentiable and $\mathcal{X} = \mathbb{R}^n$ is Theorem 2.1.8 in [8]. The proof for general strongly convex function h and general convex set $\mathcal{X}$ is in [7]. ■

B. Basic Properties

This subsection presents preliminary results related to the virtual queue update (Lemmas 3-6) that are proven for Algorithm 2 in [7].

Lemma 3 (Lemma 3 in [7]): In Algorithm 3, we have

1) At each iteration $t \in \{0, 1, 2, \ldots\}$, $Q_k(t) \geq 0$ for all $k \in \{1, 2, \ldots, m\}$.
2) At each iteration $t \in \{0, 1, 2, \ldots\}$, $Q_k(t) + g_k(\mathbf{x}(t-1)) \geq 0$ for all $k \in \{1, 2 \ldots, m\}$.
3) At iteration $t = 0$, $\|\mathbf{Q}(0)\|^2 \leq \|\mathbf{g}(\mathbf{x}(-1))\|^2$. At each iteration $t \in \{1, 2, \ldots\}$, $\|\mathbf{Q}(t)\|^2 \geq \|\mathbf{g}(\mathbf{x}(t-1))\|^2$.

Lemma 4 (Lemma 7 in [7]): Let $\mathbf{Q}(t), t \in \{0, 1, \ldots\}$ be the sequence generated by Algorithm 3. For any $t \geq 1$,

$$Q_k(t) \geq \sum_{\tau=0}^{t-1} g_k(\mathbf{x}(\tau)), \forall k \in \{1, 2, \ldots, m\}.$$

Let $\mathbf{Q}(t) = \big[Q_1(t), \ldots, Q_m(t)\big]^T$ be the vector of virtual queue backlogs. Define $L(t) = \frac{1}{2}\|\mathbf{Q}(t)\|^2$. The function $L(t)$ shall be called a *Lyapunov function*. Define the Lyapunov drift as $\Delta(t) = L(t+1) - L(t) = \frac{1}{2}[\|\mathbf{Q}(t+1)\|^2 - \|\mathbf{Q}(t)\|^2]$.

Lemma 5 (Lemma 4 in [7]): At each iteration $t \in \{0, 1, 2, \ldots\}$ in Algorithm 3, an upper bound of the Lyapunov drift is given by

$$\Delta(t) \leq \mathbf{Q}^T(t)\mathbf{g}(\mathbf{x}(t)) + \|\mathbf{g}(\mathbf{x}(t))\|^2. \tag{4}$$

Lemma 6 (Lemma 8 in [7]): Let $\mathbf{x}^*$ be an optimal solution and $\boldsymbol{\lambda}^*$ be defined in Assumption 2. Let $\mathbf{x}(t), \mathbf{Q}(t), t \in \{0, 1, \ldots\}$ be sequences generated by Algorithm 3. Then, $\sum_{\tau=0}^{t-1} f(\mathbf{x}(\tau)) \geq tf(\mathbf{x}^*) - \|\boldsymbol{\lambda}^*\|\|\mathbf{Q}(t)\|$ for all $t \geq 1$.

III. Convergence Rate Analysis of Algorithm 3

This section analyzes the convergence rate of Algorithm 3 for problem (1)-(3).

A. Upper Bounds of the Drift-Plus-Penalty Expression

Lemma 7: Let $\mathbf{x}^*$ be an optimal solution. For all $t \geq 0$ in Algorithm 3, we have $\Delta(t) + f(\mathbf{x}(t)) \leq f(\mathbf{x}^*) + \frac{1}{2\gamma}[\|\mathbf{x}^* - \mathbf{x}(t-1)\|^2 - \|\mathbf{x}^* - \mathbf{x}(t)\|^2] + \frac{1}{2}[\|\mathbf{g}(\mathbf{x}(t))\|^2 - \|\mathbf{g}(\mathbf{x}(t-1))\|^2] + \frac{1}{2}\big[\beta^2 + L_f + \|\mathbf{Q}(t)\|\|\mathbf{L_g}\| + C\|\mathbf{L_g}\| - \frac{1}{\gamma}\big]\|\mathbf{x}(t) - \mathbf{x}(t-1)\|^2$, where β, L_f, $\mathbf{L_g}$ and C are defined in Assumption 1.

Proof: Fix $t \geq 0$. Recall that $\phi(\mathbf{x}) = f(\mathbf{x}) + [\mathbf{Q}(t) + \mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{g}(\mathbf{x})$ as defined in Algorithm 3. Note that part 2 in Lemma 3 implies that $\mathbf{Q}(t) + \mathbf{g}(\mathbf{x}(t-1))$ is component-wise nonnegative. Hence, $\phi(\mathbf{x})$ is convex. Since $\mathbf{d}(t) = \nabla\phi(\mathbf{x}(t-1))$, the projection operator in Algorithm 3 can be reinterpreted as an optimization problem:

$$\begin{aligned}\mathbf{x}(t) =& \mathcal{P}_{\mathcal{X}}\left[\mathbf{x}(t-1) - \gamma\mathbf{d}(t)\right]\\ \overset{(a)}{=}& \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \Big[\phi(\mathbf{x}(t-1)) + \nabla^T\phi(\mathbf{x}(t-1))[\mathbf{x} - \mathbf{x}(t-1)]\\ &+ \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}(t-1)\|^2\Big],\end{aligned} \tag{5}$$

where (a) follows by removing the constant term $\phi(\mathbf{x}(t-1))$ in the minimization, completing the square, and using the fact that the projection of a point onto a set is equivalent to the minimization of the Euclidean distance to this point over the same set. (See [9] for the detailed proof.)

Since $\frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}(t-1)\|^2$ is strongly convex with respect to $\mathbf{x}$ with modulus $\frac{1}{\gamma}$, it follows that $\phi(\mathbf{x}(t-1)) + \nabla^T\phi(\mathbf{x}(t-1))[\mathbf{x} - \mathbf{x}(t-1)] + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}(t-1)\|^2$ is strongly convex with respect to $\mathbf{x}$ with modulus $\frac{1}{\gamma}$.

Since $\mathbf{x}(t)$ is chosen to minimize the above strongly convex function, by Lemma 2, we have

$$\begin{aligned}&\phi(\mathbf{x}(t-1)) + \nabla^T\phi(\mathbf{x}(t-1))[\mathbf{x}(t) - \mathbf{x}(t-1)]\\ &+ \frac{1}{2\gamma}\|\mathbf{x}(t) - \mathbf{x}(t-1)\|^2\\ \leq& \phi(\mathbf{x}(t-1)) + \nabla^T\phi(\mathbf{x}(t-1))[\mathbf{x}^* - \mathbf{x}(t-1)]\\ &+ \frac{1}{2\gamma}\|\mathbf{x}^* - \mathbf{x}(t-1)\|^2 - \frac{1}{2\gamma}\|\mathbf{x}^* - \mathbf{x}(t)\|^2\\ \overset{(a)}{\leq}& \phi(\mathbf{x}^*) + \frac{1}{2\gamma}[\|\mathbf{x}^* - \mathbf{x}(t-1)\|^2 - \|\mathbf{x}^* - \mathbf{x}(t)\|^2]\\ \overset{(b)}{=}& f(\mathbf{x}^*) + \underbrace{[\mathbf{Q}(t) + \mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{g}(\mathbf{x}^*)}_{\leq 0}\\ &+ \frac{1}{2\gamma}[\|\mathbf{x}^* - \mathbf{x}(t-1)\|^2 - \|\mathbf{x}^* - \mathbf{x}(t)\|^2]\\ \overset{(c)}{\leq}& f(\mathbf{x}^*) + \frac{1}{2\gamma}[\|\mathbf{x}^* - \mathbf{x}(t-1)\|^2 - \|\mathbf{x}^* - \mathbf{x}(t)\|^2],\end{aligned} \tag{6}$$

where (a) follows from the convexity of $\phi(\mathbf{x})$; (b) follows from the definition of $\phi(\mathbf{x})$; and (c) follows by using the fact that $g_k(\mathbf{x}^*) \leq 0$ and $Q_k(t) + g_k(\mathbf{x}(t-1)) \geq 0$ (i.e., part 2 in Lemma 3) for all $k \in \{1, 2, \ldots, m\}$ to eliminate the term marked by an underbrace.

Recall that $f(\mathbf{x})$ is smooth on $\mathcal{X}$ with modulus L_f by Assumption 1. By Lemma 1, we have

$$\begin{aligned}f(\mathbf{x}(t)) \leq& f(\mathbf{x}(t-1)) + \nabla^T f(\mathbf{x}(t-1))[\mathbf{x}(t) - \mathbf{x}(t-1)]\\ &+ \frac{L_f}{2}\|\mathbf{x}(t) - \mathbf{x}(t-1)\|^2.\end{aligned} \tag{7}$$

Recall that each $g_k(\mathbf{x})$ is smooth on $\mathcal{X}$ with modulus L_{g_k} by Assumption 1. Thus, $[Q_k(t)+g_k(\mathbf{x}(t-1))]g_k(\mathbf{x})$ is smooth with modulus $[Q_k(t)+g_k(\mathbf{x}(t-1))]L_{g_k}$. By Lemma 1, we have $[Q_k(t)+g_k(\mathbf{x}(t-1))]g_k(\mathbf{x}(t)) \leq [Q_k(t)+g_k(\mathbf{x}(t-1))]g_k(\mathbf{x}(t-1))+[Q_k(t)+g_k(\mathbf{x}(t-1))]\nabla^T g_k(\mathbf{x}(t-1))[\mathbf{x}(t)-\mathbf{x}(t-1)]+\frac{[Q_k(t)+g_k(\mathbf{x}(t-1))]L_{g_k}}{2}\|\mathbf{x}(t)-\mathbf{x}(t-1)\|^2$.

Summing this inequality over $k\in\{1,2,\ldots,m\}$ yields

$$\begin{aligned}
&[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{g}(\mathbf{x}(t))\\
\leq&[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{g}(\mathbf{x}(t-1))+\\
&\sum_{k=1}^{m}[Q_k(t)+g_k(\mathbf{x}(t-1))]\nabla^T g_k(\mathbf{x}(t-1))[\mathbf{x}(t)-\mathbf{x}(t-1)]\\
&+\frac{[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{L_g}}{2}\|\mathbf{x}(t)-\mathbf{x}(t-1)\|^2. \qquad (8)
\end{aligned}$$

Summing up (7) and (8) together yields

$$\begin{aligned}
&f(\mathbf{x}(t))+[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{g}(\mathbf{x}(t))\\
\overset{(a)}{\leq}&\phi(\mathbf{x}(t-1))+\nabla^T\phi(\mathbf{x}(t-1))[\mathbf{x}(t)-\mathbf{x}(t-1)]\\
&+\frac{L_f+[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{L_g}}{2}\|\mathbf{x}(t)-\mathbf{x}(t-1)\|^2, \qquad (9)
\end{aligned}$$

where (a) follows from the definition of $\phi(\mathbf{x})$.

Substituting (6) into (9) yields

$$\begin{aligned}
&f(\mathbf{x}(t))+[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{g}(\mathbf{x}(t))\\
\leq&f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(t-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(t)\|^2]+\frac{1}{2}\Big[L_f\\
&+[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{L_g}-\frac{1}{\gamma}\Big]\|\mathbf{x}(t)-\mathbf{x}(t-1)\|^2. \qquad (10)
\end{aligned}$$

Note that $\mathbf{u}_1^T\mathbf{u}_2=\frac{1}{2}[\|\mathbf{u}_1\|^2+\|\mathbf{u}_2\|^2-\|\mathbf{u}_1-\mathbf{u}_2\|^2]$ for any $\mathbf{u}_1,\mathbf{u}_2\in\mathbb{R}^m$. Thus, we have $[\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{g}(\mathbf{x}(t))=\frac{1}{2}[\|\mathbf{g}(\mathbf{x}(t-1))\|^2+\|\mathbf{g}(\mathbf{x}(t))\|^2-\|\mathbf{g}(\mathbf{x}(t-1))-\mathbf{g}(\mathbf{x}(t))\|^2]$.

Substituting this into (10) and rearranging terms yields

$$\begin{aligned}
&f(\mathbf{x}(t))+\mathbf{Q}^T(t)\mathbf{g}(\mathbf{x}(t))\\
\leq&f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(t-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(t)\|^2]+\frac{1}{2}\Big[L_f\\
&+[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{L_g}-\frac{1}{\gamma}\Big]\|\mathbf{x}(t)-\mathbf{x}(t-1)\|^2\\
&+\frac{1}{2}\|\mathbf{g}(\mathbf{x}(t-1))-\mathbf{g}(\mathbf{x}(t))\|^2-\frac{1}{2}\|\mathbf{g}(\mathbf{x}(t-1))\|^2\\
&-\frac{1}{2}\|\mathbf{g}(\mathbf{x}(t))\|^2\\
\overset{(a)}{\leq}&f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(t-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(t)\|^2]+\frac{1}{2}\Big[\beta^2+\\
&L_f+[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{L_g}-\frac{1}{\gamma}\Big]\|\mathbf{x}(t)-\mathbf{x}(t-1)\|^2\\
&-\frac{1}{2}\|\mathbf{g}(\mathbf{x}(t-1))\|^2-\frac{1}{2}\|\mathbf{g}(\mathbf{x}(t))\|^2
\end{aligned}$$

where (a) follows from $\|\mathbf{g}(\mathbf{x}(t-1))-\mathbf{g}(\mathbf{x}(t))\|\leq\beta\|\mathbf{x}(t)-\mathbf{x}(t-1)\|$, which further follows from the assumption that $\mathbf{g}(\mathbf{x})$ is Lipschitz continuous with modulus β.

Summing (4) with this inequality yields

$$\begin{aligned}
&\Delta(t)+f(\mathbf{x}(t))\\
\leq&f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(t-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(t)\|^2]\\
&+\frac{1}{2}[\|\mathbf{g}(\mathbf{x}(t))\|^2-\|\mathbf{g}(\mathbf{x}(t-1))\|^2]+\frac{1}{2}\Big[\beta^2+L_f\\
&+[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{L_g}-\frac{1}{\gamma}\Big]\|\mathbf{x}(t)-\mathbf{x}(t-1)\|^2\\
\overset{(a)}{\leq}&f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(t-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(t)\|^2]\\
&+\frac{1}{2}[\|\mathbf{g}(\mathbf{x}(t))\|^2-\|\mathbf{g}(\mathbf{x}(t-1))\|^2]+\frac{1}{2}\Big[\beta^2+L_f\\
&+\|\mathbf{Q}(t)\|\|\mathbf{L_g}\|+C\|\mathbf{L_g}\|-\frac{1}{\gamma}\Big]\|\mathbf{x}(t)-\mathbf{x}(t-1)\|^2,
\end{aligned}$$

where (a) follows from $[\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))]^T\mathbf{L_g}\leq\|\mathbf{Q}(t)+\mathbf{g}(\mathbf{x}(t-1))\|\|\mathbf{L_g}\|\leq[\|\mathbf{Q}(t)\|+\|\mathbf{g}(\mathbf{x}(t))\|]\|\mathbf{L_g}\|\leq\|\mathbf{Q}(t)\|\|\mathbf{L_g}\|+C\|\mathbf{L_g}\|$, where the first step follows from Cauchy-Schwartz inequality, the second follows from the triangular inequality and the third follows from $\|\mathbf{g}(\mathbf{x})\|\leq C$ for all $\mathbf{x}\in\mathcal{X}$, i.e., Assumption 1. ■

Lemma 8: Let $\mathbf{x}^*$ be an optimal solution and $\boldsymbol{\lambda}^*$ be defined in Assumption 2. Define $D=\beta^2+L_f+2\|\boldsymbol{\lambda}^*\|\|\mathbf{L_g}\|+2C\|\mathbf{L_g}\|$, where β, L_f, $\mathbf{L_g}$ and C are defined in Assumption 1. If $\gamma>0$ in Algorithm 3 satisfies

$$D+\|\mathbf{L_g}\|\frac{R}{\sqrt{\gamma}}-\frac{1}{\gamma}\leq 0, \qquad (11)$$

where R is defined in Assumption 1, e.g.,

$$0<\gamma\leq 1/(\|\mathbf{L_g}\|R+\sqrt{D})^2, \qquad (12)$$

then at each iteration $t\in\{0,1,2,\ldots\}$, we have

1) $\|\mathbf{Q}(t)\|\leq 2\|\boldsymbol{\lambda}^*\|+\frac{R}{\sqrt{\gamma}}+C$.
2) $\Delta(t)+f(\mathbf{x}(t))\leq f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(t-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(t)\|^2]+\frac{1}{2}[\|\mathbf{g}(\mathbf{x}(t))\|^2-\|\mathbf{g}(\mathbf{x}(t-1))\|^2]$.

Proof: Before the main proof, we verify that γ given by (12) satisfies (11). Need to choose $\gamma>0$ such that

$$\begin{aligned}
&D+\|\mathbf{L_g}\|\frac{R}{\sqrt{\gamma}}-\frac{1}{\gamma}\leq 0\Leftrightarrow D\gamma+\|\mathbf{L_g}\|R\sqrt{\gamma}-1\leq 0\\
\Leftrightarrow&\sqrt{\gamma}\leq\frac{-\|\mathbf{L_g}\|R+\sqrt{\|\mathbf{L_g}\|^2R^2+4D}}{2D}\\
&=\frac{2}{\|\mathbf{L_g}\|R+\sqrt{\|\mathbf{L_g}\|^2R^2+4D}}.
\end{aligned}$$

Note that $\frac{2}{\|\mathbf{L_g}\|R+\sqrt{\|\mathbf{L_g}\|^2R^2+4D}}\overset{(a)}{\geq}\frac{2}{2\|\mathbf{L_g}\|R+2\sqrt{D}}=\frac{1}{\|\mathbf{L_g}\|R+\sqrt{D}}$, where (a) follows from $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b},\forall a,b\geq 0$. Thus, if $\sqrt{\gamma}\leq\frac{1}{\|\mathbf{L_g}\|R+\sqrt{D}}$, i.e., $0<\gamma\leq\frac{1}{(\|\mathbf{L_g}\|R+\sqrt{D})^2}$, then inequality (11) holds. Next, we prove this lemma by induction.

- Consider $t=0$. $\|\mathbf{Q}(0)\|\leq 2\|\boldsymbol{\lambda}^*\|+\frac{R}{\sqrt{\gamma}}+C$ follows from the fact that $\|\mathbf{Q}(0)\|\overset{(a)}{\leq}\|\mathbf{g}(\mathbf{x}(-1))\|\overset{(b)}{\leq}C$, where (a) follows from part 3 in Lemma 3 and (b) follows

from Assumption 1. Thus, the first part in this lemma holds at iteration $t=0$. Note that

$$\begin{aligned}
&\beta^2+L_f+\|\mathbf{Q}(0)\|\|\mathbf{L_g}\|+C\|\mathbf{L_g}\|-1/\gamma\\
\overset{(a)}{\leq}&\beta^2+L_f+\big(2\|\boldsymbol{\lambda}^*\|+\frac{R}{\sqrt{\gamma}}+C\big)\|\mathbf{L_g}\|+C\|\mathbf{L_g}\|-\frac{1}{\gamma}\\
\overset{(b)}{=}&D+\|\mathbf{L_g}\|\frac{R}{\sqrt{\gamma}}-\frac{1}{\gamma}\overset{(c)}{\leq}0,
\end{aligned}\tag{13}$$

where (a) follows from $\|\mathbf{Q}(0)\|\leq 2\|\boldsymbol{\lambda}^*\|+\frac{R}{\sqrt{\gamma}}+C$; (b) follows from the definition of D; and (c) follows from (11), i.e., the selection rule of γ.

Applying Lemma 7 at iteration $t=0$ yields

$$\begin{aligned}
&\Delta(0)+f(\mathbf{x}(0))\\
\leq&f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(0)\|^2]\\
&+\frac{1}{2}[\|\mathbf{g}(\mathbf{x}(0))\|^2-\|\mathbf{g}(\mathbf{x}(-1))\|^2]+\frac{1}{2}\Big[\beta^2+L_f\\
&+\|\mathbf{Q}(0)\|\|\mathbf{L_g}\|+C\|\mathbf{L_g}\|-\frac{1}{\gamma}\Big]\|\mathbf{x}(0)-\mathbf{x}(-1)\|^2\\
\overset{(a)}{\leq}&f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(0)\|^2]\\
&+\frac{1}{2}[\|\mathbf{g}(\mathbf{x}(0))\|^2-\|\mathbf{g}(\mathbf{x}(-1))\|^2],
\end{aligned}$$

where (a) follows from (13). Thus, the second part in this lemma holds at iteration $t=0$.

- Assume $\Delta(\tau)+f(\mathbf{x}(\tau))\leq f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(\tau-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(\tau)\|^2]+\frac{1}{2}[\|\mathbf{g}(\mathbf{x}(\tau))\|^2-\|\mathbf{g}(\mathbf{x}(\tau-1))\|^2]$ holds for all $0\leq\tau\leq t$ and consider iteration $t+1$.

  Summing this inequality over $\tau\in\{0,1,\ldots,t\}$ yields $\sum_{\tau=0}^t\Delta(\tau)+\sum_{\tau=0}^t f(\mathbf{x}(\tau))\leq(t+1)f(\mathbf{x}^*)+\frac{1}{2\gamma}\sum_{\tau=0}^t[\|\mathbf{x}^*-\mathbf{x}(\tau-1)\|^2-\|\mathbf{x}^*-\mathbf{x}(\tau)\|^2]+\frac{1}{2}\sum_{\tau=0}^t[\|\mathbf{g}(\mathbf{x}(\tau))\|^2-\|\mathbf{g}(\mathbf{x}(\tau-1))\|^2]$.

  Recalling that $\Delta(\tau)=L(\tau+1)-L(\tau)$ and simplifying the summations yields $L(t+1)-L(0)+\sum_{\tau=0}^t f(\mathbf{x}(\tau))\leq(t+1)f(\mathbf{x}^*)+\frac{1}{2\gamma}\|\mathbf{x}^*-\mathbf{x}(-1)\|^2-\frac{1}{2\gamma}\|\mathbf{x}^*-\mathbf{x}(t)\|^2+\frac{1}{2}\|\mathbf{g}(\mathbf{x}(t))\|^2-\frac{1}{2}\|\mathbf{g}(\mathbf{x}(-1))\|^2\leq(t+1)f(\mathbf{x}^*)+\frac{1}{2\gamma}\|\mathbf{x}^*-\mathbf{x}(-1)\|^2+\frac{1}{2}\|\mathbf{g}(\mathbf{x}(t))\|^2-\frac{1}{2}\|\mathbf{g}(\mathbf{x}(-1))\|^2$. Rearranging terms yields

$$\begin{aligned}
&\sum_{\tau=0}^t f(\mathbf{x}(\tau))\\
\leq&(t+1)f(\mathbf{x}^*)+\frac{1}{2\gamma}\|\mathbf{x}^*-\mathbf{x}(-1)\|^2+\frac{1}{2}\|\mathbf{g}(\mathbf{x}(t))\|^2\\
&-\frac{1}{2}\|\mathbf{g}(\mathbf{x}(-1))\|^2+L(0)-L(t+1)\\
\overset{(a)}{=}&(t+1)f(\mathbf{x}^*)+\frac{1}{2\gamma}\|\mathbf{x}^*-\mathbf{x}(-1)\|^2+\frac{1}{2}\|\mathbf{g}(\mathbf{x}(t))\|^2\\
&-\frac{1}{2}\|\mathbf{g}(\mathbf{x}(-1))\|^2+\frac{1}{2}\|\mathbf{Q}(0)\|^2-\frac{1}{2}\|\mathbf{Q}(t+1)\|^2\\
\overset{(b)}{\leq}&(t+1)f(\mathbf{x}^*)+\frac{R^2}{2\gamma}+\frac{C^2}{2}-\frac{1}{2}\|\mathbf{Q}(t+1)\|^2,
\end{aligned}\tag{14}$$

where (a) follows from $L(0)=\frac{1}{2}\|\mathbf{Q}(0)\|^2$ and $L(t+1)=\frac{1}{2}\|\mathbf{Q}(t+1)\|^2$; (b) follows from $\|\mathbf{x}-\mathbf{y}\|\leq R$ for all $\mathbf{x},\mathbf{y}\in\mathcal{X}$, i.e., Assumption 1, $\|\mathbf{g}(\mathbf{x}(t))\|\leq C$, i.e., Assumption 1, and $\|\mathbf{Q}(0)\|^2\leq\|\mathbf{g}(\mathbf{x}(-1))\|^2$, i.e., part 3 in Lemma 3.

Applying Lemma 6 at iteration $t+1$ yields $\sum_{\tau=0}^t f(\mathbf{x}(\tau))\geq(t+1)f(\mathbf{x}^*)-\|\boldsymbol{\lambda}^*\|\|\mathbf{Q}(t+1)\|$. Combining this inequality with (14) and cancelling the common term $(t+1)f(\mathbf{x}^*)$ on both sides yields

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{Q}(t+1)\|^2-\|\boldsymbol{\lambda}^*\|\|\mathbf{Q}(t+1)\|-\frac{R^2}{2\gamma}-\frac{C^2}{2}\leq0\\
\Rightarrow&(\|\mathbf{Q}(t+1)\|-\|\boldsymbol{\lambda}^*\|)^2\leq\|\boldsymbol{\lambda}^*\|^2+\frac{R^2}{\gamma}+C^2\\
\Rightarrow&\|\mathbf{Q}(t+1)\|\leq\|\boldsymbol{\lambda}^*\|+\sqrt{\|\boldsymbol{\lambda}^*\|^2+R^2/\gamma+C^2}\\
\overset{(a)}{\Rightarrow}&\|\mathbf{Q}(t+1)\|\leq2\|\boldsymbol{\lambda}^*\|+R/\sqrt{\gamma}+C,
\end{aligned}$$

where (a) follows from the basic inequality $\sqrt{a+b+c}\leq\sqrt{a}+\sqrt{b}+\sqrt{c}$ for any $a,b,c\geq0$. Thus, the first part in this lemma holds at iteration $t+1$.

Note that

$$\begin{aligned}
&\beta^2+L_f+\|\mathbf{Q}(t+1)\|\|\mathbf{L_g}\|+C\|\mathbf{L_g}\|-\frac{1}{\gamma}\\
\overset{(a)}{\leq}&\beta^2+L_f+(2\|\boldsymbol{\lambda}^*\|+\frac{R}{\sqrt{\gamma}}+C)\|\mathbf{L_g}\|+C\|\mathbf{L_g}\|-\frac{1}{\gamma}\\
\overset{(b)}{=}&D+\|\mathbf{L_g}\|\frac{R}{\sqrt{\gamma}}-\frac{1}{\gamma}\overset{(c)}{\leq}0,
\end{aligned}\tag{15}$$

where (a) follows from $\|\mathbf{Q}(t+1)\|\leq2\|\boldsymbol{\lambda}^*\|+\frac{R}{\sqrt{\gamma}}+C$; (b) follows from the definition of D; and (c) follows from (11), i.e., the selection rule of γ.

Applying Lemma 7 at iteration $t+1$ yields

$$\begin{aligned}
&\Delta(t+1)+f(\mathbf{x}(t+1))\\
\leq&f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(t)\|^2-\|\mathbf{x}^*-\mathbf{x}(t+1)\|^2]\\
&+\frac{1}{2}[\|\mathbf{g}(\mathbf{x}(t+1))\|^2-\|\mathbf{g}(\mathbf{x}(t))\|^2]+\frac{1}{2}\Big[\beta^2+L_f+\\
&\|\mathbf{Q}(t+1)\|\|\mathbf{L_g}\|+C\|\mathbf{L_g}\|-\frac{1}{\gamma}\Big]\|\mathbf{x}(t+1)-\mathbf{x}(t)\|^2\\
\overset{(a)}{\leq}&f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(t)\|^2-\|\mathbf{x}^*-\mathbf{x}(t+1)\|^2]\\
&+\frac{1}{2}[\|\mathbf{g}(\mathbf{x}(t+1))\|^2-\|\mathbf{g}(\mathbf{x}(t))\|^2],
\end{aligned}$$

where (a) follows from (15). Thus, the second part in this lemma holds at iteration $t+1$.

Thus, both parts in this lemma follow by induction. ■

Remark 1: Recall that if each $g_k(\mathbf{x})$ is a linear function, then $L_{g_k}=0$ for all $k\in\{1,2,\ldots,m\}$. In this case, equation (12) reduces to $0<\gamma\leq1/(\beta^2+L_f)$.

B. Objective Value Violations

Theorem 1 (Objective Value Violations): Let $\mathbf{x}^*$ be an optimal solution. If we choose γ according to (12) in Algorithm 3, then for all $t\geq1$, we have $f(\overline{\mathbf{x}}(t))\leq f(\mathbf{x}^*)+\frac{1}{t}\frac{R^2}{2\gamma}$, where R is defined in Assumption 1.

Proof: Fix $t\geq1$. By part 2 in Lemma 8, we have $\Delta(\tau)+f(\mathbf{x}(\tau))\leq f(\mathbf{x}^*)+\frac{1}{2\gamma}[\|\mathbf{x}^*-\mathbf{x}(\tau-1)\|^2-$

$\|\mathbf{x}^* - \mathbf{x}(\tau)\|] + \frac{1}{2}[\|\mathbf{g}(\mathbf{x}(\tau))\|^2 - \|\mathbf{g}(\mathbf{x}(\tau-1))\|^2]$ for all $\tau \in \{0,1,2,\ldots\}$.

Summing over $\tau \in \{0,1,\ldots,t-1\}$ yields $\sum_{\tau=0}^{t-1}\Delta(\tau) + \sum_{\tau=0}^{t-1} f(\mathbf{x}(\tau)) \leq tf(\mathbf{x}^*) + \frac{1}{2\gamma}\sum_{\tau=0}^{t-1}[\|\mathbf{x}^* - \mathbf{x}(\tau-1)\|^2 - \|\mathbf{x}^* - \mathbf{x}(\tau)\|^2] + \frac{1}{2}\sum_{\tau=0}^{t-1}[\|\mathbf{g}(\mathbf{x}(\tau))\|^2 - \|\mathbf{g}(\mathbf{x}(\tau-1))\|^2]$.

Recalling that $\Delta(\tau) = L(\tau+1) - L(\tau)$ and simplifying the summations yields $L(t) - L(0) + \sum_{\tau=0}^{t-1} f(\mathbf{x}(\tau)) \leq tf(\mathbf{x}^*) + \frac{1}{2\gamma}\|\mathbf{x}^* - \mathbf{x}(-1)\|^2 - \frac{1}{2\gamma}\|\mathbf{x}^* - \mathbf{x}(t-1)\|^2 + \frac{1}{2}\|\mathbf{g}(\mathbf{x}(t-1))\|^2 - \frac{1}{2}\|\mathbf{g}(\mathbf{x}(-1))\|^2 \leq tf(\mathbf{x}^*) + \frac{1}{2\gamma}\|\mathbf{x}^* - \mathbf{x}(-1)\|^2 + \frac{1}{2}\|\mathbf{g}(\mathbf{x}(t-1))\|^2 - \frac{1}{2}\|\mathbf{g}(\mathbf{x}(-1))\|^2$. Rearranging terms yields

$$\begin{aligned}
&\sum_{\tau=0}^{t-1} f(\mathbf{x}(\tau)) \\
\leq& tf(\mathbf{x}^*) + \frac{1}{2\gamma}\|\mathbf{x}^* - \mathbf{x}(-1)\|^2 + \frac{1}{2}\|\mathbf{g}(\mathbf{x}(t-1))\|^2 \\
&- \frac{1}{2}\|\mathbf{g}(\mathbf{x}(-1))\|^2 + L(0) - L(t) \\
\overset{(a)}{=}& tf(\mathbf{x}^*) + \frac{1}{2\gamma}\|\mathbf{x}^* - \mathbf{x}(-1)\|^2 + \frac{1}{2}\|\mathbf{g}(\mathbf{x}(t-1))\|^2 \\
&- \frac{1}{2}\|\mathbf{g}(\mathbf{x}(-1))\|^2 + \frac{1}{2}\|\mathbf{Q}(0)\|^2 - \frac{1}{2}\|\mathbf{Q}(t)\|^2 \\
\overset{(b)}{\leq}& tf(\mathbf{x}^*) + \frac{1}{2\gamma}\|\mathbf{x}^* - \mathbf{x}(-1)\|^2 \overset{(c)}{\leq} tf(\mathbf{x}^*) + \frac{R^2}{2\gamma},
\end{aligned}$$

where (a) follows from the definition that $L(0) = \frac{1}{2}\|\mathbf{Q}(0)\|^2$ and $L(t) = \frac{1}{2}\|\mathbf{Q}(t)\|^2$; (b) follows from the fact that $\|\mathbf{Q}(0)\|^2 \leq \|\mathbf{g}(\mathbf{x}(-1))\|^2$ and $\|\mathbf{Q}(t)\|^2 \geq \|\mathbf{g}(\mathbf{x}(t-1))\|^2$ for $t \geq 1$, i.e., part 3 in Lemma 3; and (c) follows from the fact that $\|\mathbf{x} - \mathbf{y}\| \leq R$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, i.e., Assumption 1.

Dividing both sides by factor t yields $\frac{1}{t}\sum_{\tau=0}^{t-1} f(\mathbf{x}(\tau)) \leq f(\mathbf{x}^*) + \frac{1}{t}\frac{R^2}{2\gamma}$. Finally, since $\overline{\mathbf{x}}(t) = \frac{1}{t}\sum_{\tau=0}^{t-1}\mathbf{x}(\tau)$ and $f(\mathbf{x})$ is convex, By Jensen's inequality it follows that $f(\overline{\mathbf{x}}(t)) \leq \frac{1}{t}\sum_{\tau=0}^{t-1} f(\mathbf{x}(\tau))$. ∎

C. Constraint Violations

Theorem 2 (Constraint Violations): Let $\mathbf{x}^*$ be an optimal solution and $\boldsymbol{\lambda}^*$ be defined in Assumption 2. If we choose γ according to (12) in Algorithm 3, then for all $t \geq 1$, the constraints satisfy $g_k(\overline{\mathbf{x}}(t)) \leq \frac{1}{t}\big(2\|\boldsymbol{\lambda}^*\| + \frac{R}{\sqrt{\gamma}} + C\big), \forall k \in \{1,2,\ldots,m\}$, where R and C are defined in Assumption 1.

Proof: Fix $t \geq 1$ and $k \in \{1,2,\ldots,m\}$. Recall that $\overline{\mathbf{x}}(t) = \frac{1}{t}\sum_{\tau=0}^{t-1}\mathbf{x}(\tau)$. Thus,

$$\begin{aligned}
g_k(\overline{\mathbf{x}}(t)) &\overset{(a)}{\leq} \frac{1}{t}\sum_{\tau=0}^{t-1} g_k(\mathbf{x}(\tau)) \overset{(b)}{\leq} \frac{Q_k(t)}{t} \leq \frac{\|\mathbf{Q}(t)\|}{t} \\
&\overset{(c)}{\leq} \frac{1}{t}\big(2\|\boldsymbol{\lambda}^*\| + \frac{R}{\sqrt{\gamma}} + C\big),
\end{aligned}$$

where (a) follows from the convexity of $g_k(\mathbf{x})$ and Jensen's inequality; (b) follows from Lemma 4; and (c) follows from part 1 in Lemma 8. ∎

Theorems 1 and 2 show that Algorithm 3 ensures error decays like $O(1/t)$ and provides an ϵ-approximiate solution with convergence time $O(1/\epsilon)$.

D. Practical Implementations

By Theorems 1 and 2, it suffices to choose γ according to (12) to guarantee the $O(1/t)$ convergence rate of Algorithm 3. If all constraint functions are linear, then (12) is independent of $\|\boldsymbol{\lambda}^*\|$ by Remark 1. For general constraint functions, we need to know the value of $\|\boldsymbol{\lambda}^*\|$, which is typically unknown, to select γ according to (12). However, it is easy to observe that an upper bound of $\|\boldsymbol{\lambda}^*\|$ is sufficient for us to choose γ satisfying (12). To obtain an upper bound of $\|\boldsymbol{\lambda}^*\|$, the next lemma is useful if problem (1)-(3) has an interior feasible point, i.e., the Slater condition is satisfied.

Lemma 9 (Lemma 1 in [5]): Consider convex program

$$\begin{aligned}
\min \quad & f(\mathbf{x}) \\
\text{s.t.} \quad & g_k(\mathbf{x}) \leq 0, k \in \{1,2,\ldots,m\} \\
& \mathbf{x} \in \mathcal{X} \subseteq R^n
\end{aligned}$$

and define the Lagrangian dual function as $q(\boldsymbol{\lambda}) = \inf_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{g}(\mathbf{x})\}$. If the Slater condition holds, i.e., there exists $\hat{\mathbf{x}} \in X$ such that $g_j(\mathbf{x}) < 0, \forall j \in \{1,2,\ldots,m\}$, then the level sets $\mathcal{V}_{\hat{\boldsymbol{\lambda}}} = \{\boldsymbol{\lambda} \geq \mathbf{0} : q(\boldsymbol{\lambda}) \geq q(\hat{\boldsymbol{\lambda}})\}$ are bounded for any $\hat{\boldsymbol{\lambda}}$. In particular, we have $\max_{\boldsymbol{\lambda}\in\mathcal{V}_{\hat{\boldsymbol{\lambda}}}}\|\boldsymbol{\lambda}\| \leq \frac{1}{\min_{1\leq j\leq m}\{-g_j(\hat{\mathbf{x}})\}}(f(\hat{\mathbf{x}}) - q(\hat{\boldsymbol{\lambda}}))$.

By Lemma 9, if convex program (1)-(3) has a feasible point $\hat{\mathbf{x}} \in \mathcal{X}$ such that $g_k(\hat{\mathbf{x}}) < 0, \forall k \in \{1,2,\ldots,m\}$, then we can take an arbitrary $\hat{\boldsymbol{\lambda}} \geq \mathbf{0}$ to obtain the value $q(\hat{\boldsymbol{\lambda}}) = \inf_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x}) + \hat{\boldsymbol{\lambda}}^T\mathbf{g}(\mathbf{x})\}$ and conclude that $\|\boldsymbol{\lambda}^*\| \leq \frac{1}{\min_{1\leq j\leq m}\{-g_j(\hat{\mathbf{x}})\}}(f(\hat{\mathbf{x}}) - q(\hat{\boldsymbol{\lambda}}))$. Since $f(\mathbf{x})$ is continuous and $\mathcal{X}$ is a compact set, there exists constant $F > 0$ such that $|f(\mathbf{x})| \leq F$ for all $\mathbf{x} \in \mathcal{X}$. Thus, we can take $\hat{\boldsymbol{\lambda}} = \mathbf{0}$ such that $q(\hat{\boldsymbol{\lambda}}) = \min_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x})\} \geq -F$. It follows from Lemma 9 that $\|\boldsymbol{\lambda}^*\| \leq \frac{1}{\min_{1\leq j\leq m}\{-g_j(\hat{\mathbf{x}})\}}(f(\hat{\mathbf{x}}) - q(\hat{\boldsymbol{\lambda}})) \leq \frac{2F}{\min_{1\leq j\leq m}\{-g_j(\hat{\mathbf{x}})\}}$.

REFERENCES

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[2] 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.

[3] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.

[4] M. J. Neely, "Distributed and secure computation of convex programs over a network of connected processors," in *DCDIS Conference Guelph*, July 2005.

[5] A. Nedić and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1757–1780, 2009.

[6] M. J. Neely, "A simple convergence time analysis of drift-plus-penalty for stochastic optimization and convex programs," *arXiv:1412.0791*, 2014.

[7] H. Yu and M. J. Neely, "A simple parallel algorithm with an $O(1/t)$ convergence rate for general convex programs," *arXiv:1512.08370*, 2015.

[8] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2004.

[9] H. Yu and M. J. Neely, "A primal-dual type algorithm with the $O(1/t)$ convergence rate for large scale constrained convex programs," *arXiv:1604.02216*, 2016.