Time-Average Optimization with Non-Convex Decision Set and Its Convergence

Sucha Supittayapornpong, Longbo Huang, Michael J. Neely

***Abstract*—This paper considers *time-average optimization*, where a decision vector is chosen every time step within a (possibly non-convex) set, and the goal is to minimize a convex function of the time averages subject to convex constraints on these averages. Such problems have applications in networking, multi-agent systems, and operations research, where decisions are constrained to a discrete set and the decision average can represent average bit rates or average agent actions. This time-average optimization extends traditional convex formulations to allow a non-convex decision set. This class of problems can be solved by Lyapunov optimization. A simple drift-based algorithm, related to a classical dual subgradient algorithm, converges to an ϵ-optimal solution within $O(1/\epsilon^2)$ time steps. Further, the algorithm is shown to have a transient phase and a steady state phase which can be exploited to improve convergence rates to $O(1/\epsilon)$ and $O(1/\epsilon^{1.5})$ when vectors of Lagrange multipliers satisfy locally-polyhedral and locally-smooth assumptions respectively. Practically, this improved convergence suggests that decisions should be implemented after the transient period.**

I. INTRODUCTION

Convex optimization is often used to optimally control communication networks (see [1] and references therein) and distributed multi-agent systems [2]. This framework utilizes both convexity properties of an objective function and a feasible decision set. However, various systems have inherent discrete (and hence non-convex) decision sets. For example, a wireless system might constrain transmission rates to a finite set corresponding to a fixed set of coding options. Further, distributed agents might only have finite options of decisions. This discreteness restrains the application of convex optimization.

Let I and J be positive integers. This paper considers a class of problems called *time-average optimization* where decision vectors $x(t) = (x_1(t), \ldots, x_I(t))$ are chosen sequentially over time slots $t \in \{0, 1, 2, \ldots\}$ from a decision set $\mathcal{X}$, which is a closed and bounded subset of $\mathbb{R}^I$ (possibly non-convex and discrete), and its average $\bar{x} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} x(t)$ solves the following problem:

$$\begin{aligned} \text{Minimize} \quad & f(\bar{x}) && (1)\\ \text{Subject to} \quad & g_j(\bar{x}) \leq 0 && j \in \{1, \ldots, J\} \\ & x(t) \in \mathcal{X} && t \in \{0, 1, 2, \ldots\}, \end{aligned}$$

where $f : \overline{\mathcal{X}} \to \mathbb{R}$ and $g_j : \overline{\mathcal{X}} \to \mathbb{R}$ are convex and continuous functions and $\overline{\mathcal{X}}$ is the convex hull of $\mathcal{X}$.

This time-average optimization reflects a scenario where an objective is in the time-average sense. For example, network users are interested in average bit rates or throughput, and distributed agents are concerned with average actions. The formulation can be considered as a fine granularity version of a one-shot average formulation, where an average decision is chosen, and can be used to extend several convex optimization problems in the literature, see for example [1] and references therein, to have non-convex decision sets.

Formulation (1) has an optimal solution which can be converted (by averaging) to the following convex optimization problem:

$$\begin{aligned} \text{Minimize} \quad & f(x) && (2)\\ \text{Subject to} \quad & g_j(x) \leq 0 && j \in \{1, \ldots, J\} \\ & x \in \overline{\mathcal{X}}. \end{aligned}$$

Note that an optimal solution to formulation (2) may not be in the non-convex decision set $\mathcal{X}$. Nevertheless, problems (1) and (2) have the same optimal value. In addition, directly applying a primal-average technique in [3], which can be traced back to the work in [4], on a non-convex formation where the convex hull in (2) is removed may lead to a local optimal solution with respect to the time-average problem (1). For example, when $\mathcal{X} = \{0, 1\}, J = 1, f(x) = (x - 2/3)^2, g_1(x) = 2/3 - x$, a primal average solution from the technique in [3] is 1, while a solution to problem (1) is $\bar{x} = 2/3$.

Although there have been several techniques utilizing time-average solutions [3], [5], [6], those works are limited to convex formulations. In fact, this work can be considered as a generalization of [3], [6] as decisions are allowed to be chosen from a non-convex set. A non-convex optimization problem is considered in [7], where an approximate problem is solved with the assumption of a unique vector of Lagrange multipliers. In comparison, when $f(x)$ and $g_j(x)$'s are Lipschitz continuous, the basic algorithm proposed in this paper solves problem (1) without the uniqueness assumption. However, the uniqueness assumption is used to prove faster convergence time for the refined algorithms of this paper. This paper is inspired by the Lyapunov optimization technique [8] which solves stochastic and time-average optimization

This material is supported in part by one or more of: the NSF Career grant CCF-0747525, the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory W911NF-09-2-0053. The work of Longbo Huang was supported in part by the National Natural Science Foundation of China Grants 61672316, 61303195, the Tsinghua Initiative Research Grant, and the China youth 1000-talent grant.

S. Supittayapornpong and M. J. Neely are with Electrical Engineering Department, University of Southern California, 3740 McClintock Ave., Los Angeles, CA, USA 90089-2565, Tel: +1-213-740-4685, Fax: +1-213-740-8729 supittay@usc.edu, mjneely@usc.edu

L. Huang is with Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China, 100084, Tel: +8610-62781693, Fax: +8610-62797331-2000 longbohuang@tsinghua.edu.cn

The corresponding author is Mr. Sucha Supittayapornpong.

problems, including problems such as (1). This paper forms the connection between the technique and a general convex optimization to analyze a convergence time of a *drift-plus-penalty* algorithm, which solves problem (1). Importantly, this paper shows that faster convergence can be achieved by starting time averages after a suitable period.

Another area of literature focuses on convergence time of first-order algorithms to an ϵ-optimal solution to a convex problem, including problem (2). For *unconstrained optimization* without strong convexity of the objective function, the accelerated method (with Lipschitz continuous gradients) has $O(1/\sqrt{\epsilon})$ convergence time [9], [10], while gradient and subgradient methods take $O(1/\epsilon)$ and $O(1/\epsilon^2)$ respectively [3], [11]. Two $O(1/\epsilon)$ first-order methods for *constrained optimization* are developed in [12], [13], but the results rely on special convex formulations. A second-order method for constrained optimization [14] has a fast convergence rate but relies on special a convex formulation. All of these results rely on convexity assumptions that do not hold in formulation (1).

This paper develops an algorithm for the formulation (1) and analyzes its convergence time. The algorithm is shown to have $O(1/\epsilon^2)$ convergence time with a mild Slater condition. However, inspired by results in [15], under a uniqueness assumption on Lagrange multipliers the algorithm is shown to enter two phases: a *transient phase* and a *steady state phase*. Convergence time can be significantly improved by starting the time averages after the transient phase. Specifically, when a dual function satisfies a *locally-polyhedral* assumption, the modified algorithm has $O(1/\epsilon)$ convergence time (including the time spent in the transient phase), which equals the best known convergence time for constrained convex optimization via first-order methods. On the other hand, when the dual function satisfies a *locally-smooth* assumption, the algorithm has $O(1/\epsilon^{1.5})$ convergence time. An application of these improved convergence times can be effective implementation when decisions are implemented online after offline calculation during a transient period.

The contributions of this paper are summarized below.

1) We establish the connection between Lyapunov optimization and a dual subgradient algorithm for a problem with a non-convex decision set.
2) We generalize the modeling of a one-shot convex optimization (2), extensively used in [1], to the time-average formulation (1) that allows a non-convex decision set, while optimality and complexity are preserved.
3) We investigate transient and steady-state behaviors of the algorithm solving the time-average problem (1). Then, we exploit the behaviors to obtain sequences of decisions that achieve $O(\epsilon)$-optimal solutions within $O(1/\epsilon)$ and $O(1/\epsilon^{1.5})$ iterations under locally-polyhedral and locally-smooth assumptions instead of the standard $O(1/\epsilon^2)$ iterations in [3], [6].

The paper is organized as follows. Section II constructs an algorithm to solve the time-average problem. The general $O(1/\epsilon^2)$ convergence time is proven in Section III. Section IV explores faster convergence times of $O(1/\epsilon)$ and $O(1/\epsilon^{1.5})$ under the unique Lagrange multiplier assumption. Example problems are given in Section V.

II. TIME-AVERAGE OPTIMIZATION

In order to solve problem (1), an embedded problem with a similar solution is formulated with the following assumptions.

A. The extended set $\mathcal{Y}$

Let $\mathcal{Y}$ be a closed, bounded, and convex subset of $\mathbb{R}^I$ that contains $\overline{\mathcal{X}}$. Assume the functions $f(x)$, $g_j(x)$ for $j \in \{1, \ldots, J\}$ extend as real-valued continuous and convex functions over $x \in \mathcal{Y}$. The set $\mathcal{Y}$ can be defined as $\overline{\mathcal{X}}$ itself. However, choosing $\mathcal{Y}$ as a larger set helps to ensure a Slater condition is satisfied (defined below) and simplifies the resulting optimization. For example, set $\mathcal{Y}$ might be chosen as a closed and bounded hyper-rectangle that contains $\overline{\mathcal{X}}$ in its interior.

B. Lipschitz continuity and Slater condition

In addition to assuming that $f(x)$ and $g_j(x)$ are convex over $x \in \mathcal{Y}$, assume they are *Lipschitz continuous*, so there is a constant $M > 0$ such that for all $x, y \in \mathcal{Y}$:

$$|f(x) - f(y)| \leq M\|x - y\| \tag{3}$$

$$|g_j(x) - g_j(y)| \leq M\|x - y\| \tag{4}$$

where $\|x\| = \sqrt{x_1^2 + \cdots + x_I^2}$ is the Euclidean norm.

Further, assume that there exists a vector $\hat{x} \in \overline{\mathcal{X}}$ that satisfies $g_j(\hat{x}) < 0$ for all $j \in \{1, \ldots, J\}$, and is such that $\hat{x}$ is in the interior of set $\mathcal{Y}$. This is a *Slater condition* that ensures the constraints are feasible for the problem of interest.

C. Relation to dual subgradient algorithm

Problem (1) can be solved by the Lyapunov optimization technique [8], which is identical to a classic dual subgradient method [3], [16] that solves problem (5), with the exception that it takes a time average of primal values [15], [17].

$$\begin{aligned} \text{Minimize} \quad & f(y) && \\ \text{Subject to} \quad & g_j(y) \leq 0 && j \in \{1, \ldots, J\} \\ & x_i = y_i && i \in \{1, \ldots, I\} \\ & x \in \overline{\mathcal{X}}, \quad y \in \mathcal{Y}. && \end{aligned} \tag{5}$$

Problem (5) is called the *embedded formulation* of the time-average problem (1) and is convex. It is not difficult to show that the above problem has an optimal value $f^{(\text{opt})}$ that is the same as that of problems (1) and (2). Compared to a formulation in [3], problem (5) contains additional equality constraints and a set constraint $\overline{\mathcal{X}}$. This is different from [3], whose results cannot be applied directly.

Now consider the dual of embedded formulation (5). Let vectors w and z be dual variables of the first and second constraints in problem (5), where the feasible set of (w, z) is denoted by $\Pi = \mathbb{R}_+^J \times \mathbb{R}^I$. Let $g(y) = (g_1(y), \ldots, g_J(y))$ denote a J-dimensional column vector of functions $g_j(y)$. A Lagrangian is $\Lambda(x, y, w, z) = f(y) + w^\top g(y) + z^\top(x - y)$. Define:

$$x^*(z) = \operatorname*{arginf}_{x \in \overline{\mathcal{X}}} z^\top x \quad (\text{with } x^*(z) \in \mathcal{X})$$

$$y^*(w, z) = \operatorname*{arginf}_{y \in \mathcal{Y}} [f(y) + w^\top g(y) - z^\top y].$$

Notice that $x^*(z)$ may have multiple candidates including extreme point solutions, since $z^\top x$ is a linear function. We restrict $x^*(z)$ to any of these extreme solutions, which implies $x^*(z) \in \mathcal{X}$. Then the dual function is defined as

$$\begin{aligned} d(w,z) &= \inf_{x\in\overline{\mathcal{X}}, y\in\mathcal{Y}} \Lambda(x,y,w,z) \qquad (6)\\ &= f(y^*(w,z)) + w^\top g(y^*(w,z)) + z^\top[x^*(z) - y^*(w,z)]. \end{aligned}$$

A pair of subgradients [16] with respect to w and z is:

$$\partial_w d(w,z) = g(y^*(w,z)), \quad \partial_z d(w,z) = x^*(z) - y^*(w,z).$$

Finally, the dual formulation of embedded problem (5) is

$$\text{Maximize} \quad d(w,z) \qquad \text{Subject to} \quad (w,z) \in \Pi. \qquad (7)$$

Let the optimal value of problem (7) be d^*. Since problem (5) is convex, the duality gap is zero, and $d^* = f^{(\text{opt})}$. Problem (7) can be treated by a dual subgradient method [16] with a fixed stepsize $1/V$ and the restriction on $x(t) \in \mathcal{X}$, where $V > 0$ is a paremeter. This leads to Algorithm 1 summarized in the figure below, called the *dual subgradient algorithm*. Define the operator $[x]_+$ as a projection of x onto the non-negative orthant. Note that the algorithm is different from the one in [3] due to the equality constraints and the restriction on $x(t)$.

Initialize $w(0)$ and $z(0)$.
for $t = 0, 1, 2, \ldots$ **do**
  $x(t) = \text{arginf}_{x\in\overline{\mathcal{X}}}\, z(t)^\top x$ (with $x(t) \in \mathcal{X}$)
  $y(t) = \text{arginf}_{y\in\mathcal{Y}}[f(y) + w(t)^\top g(y) - z(t)^\top y]$
  $w(t+1) = \left[w(t) + \frac{1}{V} g(y(t))\right]_+$
  $z(t+1) = z(t) + \frac{1}{V}[x(t) - y(t)]$
end

Algorithm 1: Dual subgradient algorithm with restriction

Indeed, the primal vectors $x(t)$ and $y(t)$ *do not converge* to anything near a solution in many cases, such as when the $f(x)$ and $g_j(x)$ functions are linear or piecewise linear. However, Algorithm 1 ensures that the *time averages* of $x(t)$ and $y(t)$ converge as desired.

We use the notation $w(t)$ and $z(t)$ from Algorithm 1, with the update rule for $w(t+1)$ and $z(t+1)$ given there:

$$w(t+1) = \left[w(t) + \frac{1}{V} g(y(t))\right]_+ \qquad (8)$$

$$z(t+1) = z(t) + \frac{1}{V}[x(t) - y(t)]. \qquad (9)$$

For ease of notation, define $\lambda(t) \triangleq (w(t), z(t))$ as a concatenation of these vectors. Let C be some positive constant such that $\|g(y)\|^2 \leq C$ and $\|x - y\|^2 \leq C$ for any $x \in \overline{\mathcal{X}}$ and any $y \in \mathcal{Y}$, since $\overline{\mathcal{X}}$ is closed and bounded. We first provide some useful properties. It holds that

$$\|\lambda(t+1) - \lambda(t)\| \leq \sqrt{2C}/V \quad \text{for all } t, \qquad (10)$$

since

$$\begin{aligned} &\|\lambda(t+1) - \lambda(t)\|^2 \\ &\quad = \|w(t+1) - w(t)\|^2 + \|z(t+1) - z(t)\|^2 \\ &\quad \leq \frac{1}{V^2}\|g(y(t))\|^2 + \frac{1}{V^2}\|x(t) - y(t)\|^2 \leq 2C/V^2 \qquad (11) \end{aligned}$$

where (11) follows from (8)–(9) and the definition of C.

$$\begin{aligned} &\|\lambda(t+1)\|^2 - \|\lambda(t)\|^2 \\ &\quad = \|w(t+1)\|^2 + \|z(t+1)\|^2 - \|w(t)\|^2 - \|z(t)\|^2 \\ &\quad \leq \frac{2C}{V^2} + \frac{2}{V} w(t)^\top g(y(t)) + \frac{2}{V} z(t)[x(t) - y(t)], \end{aligned}$$

where the last inequality uses the result of expanding the squared norms of (8)–(9). Since Algorithm 1 chooses $x(t)$, $y(t)$ to minimize $\Lambda(x(t), y(t), w(t), z(t))$ in (6), the above bound implies that

$$\begin{aligned} d(\lambda(t)) &= f(y(t)) + w(t)^\top g(y(t)) + z(t)^\top[x(t) - y(t)] \\ &\geq f(y(t)) + \frac{V}{2}\Big[\|\lambda(t+1)\|^2 - \|\lambda(t)\|^2\Big] - \frac{C}{V}. \qquad (12) \end{aligned}$$

From convex analysis, the dual function $d(\lambda)$, defined in (6), has the following properties [16]:

- $d(\lambda) \leq f^{(\text{opt})}$ for all $\lambda \in \Pi$.
- If the Slater condition holds, then there are real numbers $F > 0$, $\eta > 0$ such that: $d(\lambda) \leq F - \eta\|\lambda\|$ for all $\lambda \in \Pi$.
- If the Slater condition holds, then there is an optimal value $\lambda^* \in \Pi$, called a *Lagrange multiplier vector* [16], that maximizes $d(\lambda)$. Specifically, $d(\lambda^*) = f^{(\text{opt})}$.

The first two properties can be substituted into the inequality (12) to ensure that, under Algorithm 1, the following inequalities hold for all time slots $t \in \{0, 1, 2, \ldots\}$:

$$\frac{V}{2}\Big[\|\lambda(t+1)\|^2 - \|\lambda(t)\|^2\Big] + f(y(t)) \leq \frac{C}{V} + f^{(\text{opt})} \qquad (13)$$

$$\frac{V}{2}\Big[\|\lambda(t+1)\|^2 - \|\lambda(t)\|^2\Big] + f(y(t)) \leq \frac{C}{V} + F - \eta\|\lambda(t)\| \qquad (14)$$

III. GENERAL CONVERGENCE RESULT

Define the average of variables $\{a(t)\}_{t=0}^{T-1}$ as

$$\bar{a}(T) \triangleq \frac{1}{T}\sum_{t=0}^{T-1} a(t) \qquad \text{for } T \in \{1, 2, \ldots\}.$$

Theorem 1: Let $\{x(t), w(t), z(t)\}_{t=0}^{\infty}$ be a sequence generated by Algorithm 1. For $T > 0$, we have

$$\begin{aligned} f(\bar{x}(T)) - f^{(\text{opt})} &\leq \frac{V}{2T}\Big[\|\lambda(0)\|^2 - \|\lambda(T)\|^2\Big] + \frac{C}{V} \\ &\quad + \frac{VM}{T}\|z(T) - z(0)\| \qquad (15) \\ g_j(\bar{x}(T)) &\leq \frac{V}{T}|w_j(T) - w_j(0)| + \frac{VM}{T}\|z(T) - z(0)\| \\ &\quad j \in \{1, \ldots, J\}, \qquad (16) \end{aligned}$$

where M is the Lipschitz constant from (3)–(4).

Proof: For the first part, we have from the Lipschitz property (3):

$$f(\bar{x}(T)) - f^{(\text{opt})} \leq [f(\bar{y}(T)) - f^{(\text{opt})}] + M\|\bar{y}(T) - \bar{x}(T)\|. \qquad (17)$$

We first upper bound $f(\bar{y}(T)) - f^{(\text{opt})}$ on the right-hand side of (17). Let $\{x(t), y(t), w(t), z(t)\}_{t=0}^{\infty}$ be a sequence generated by Algorithm 1. Relation (13) can be rewritten as

$$f(y(t)) - f^{(\text{opt})} \leq \frac{C}{V} + \frac{V}{2}\Big[\|\lambda(t)\|^2 - \|\lambda(t+1)\|^2\Big].$$

Summing from $t = 0, \ldots, T-1$ and dividing by T gives:

$$\frac{1}{T}\sum_{t=0}^{T-1} f(y(t)) - f^{(\text{opt})} \leq \frac{C}{V} + \frac{V}{2T}\Big[\|\lambda(0)\|^2 - \|\lambda(T)\|^2\Big].$$

Using Jensen's inequality and the convexity of f gives:

$$f(\bar{y}(T)) - f^{(\text{opt})} \leq \frac{V}{2T}\Big[\|\lambda(0)\|^2 - \|\lambda(T)\|^2\Big] + \frac{C}{V}. \quad (18)$$

For $\|\bar{y}(T) - \bar{x}(T)\|$ in (17), we consider the update equation of $z(t)$ in (9). Summing from $t = 0, \ldots, T-1$ yields $z_i(T) - z_i(0) = \frac{1}{V}\sum_{t=0}^{T-1}[x_i(t) - y_i(t)]$ for every i. Rearranging and dividing by T gives:

$$\bar{x}_i(T) - \bar{y}_i(T) = \frac{V}{T}[z_i(T) - z_i(0)] \quad i \in \{1, \ldots, I\}. \quad (19)$$

Substituting (18) and (19) into (17) proves (15).

For the second part, we have from (4):

$$g_j(\bar{x}(T)) \leq g_j(\bar{y}(T)) + M\|\bar{y}(T) - \bar{x}(T)\|. \quad (20)$$

We first bound $g_j(\bar{y}(T))$. The update equation of $w(t)$ in (8) implies, for every j, that

$$w_j(t+1) = [w_j(t) + \frac{1}{V}g_j(y(t))]_+ \geq w_j(t) + \frac{1}{V}g_j(y(t)),$$

and $w_j(t+1) - w_j(t) \geq \frac{1}{V}g_j(y(t))$. Summing from $t = 0, \ldots, T-1$ yields $w_j(T) - w_j(0) \geq \frac{1}{V}\sum_{t=0}^{T-1} g_j(y(t))$. Dividing by T and using Jensen's inequality and convexity of g_j gives:

$$\frac{1}{T}[w_j(T) - w_j(0)] \geq \frac{1}{VT}\sum_{t=0}^{T-1} g_j(y(t)) \geq \frac{1}{V}g_j(\bar{y}(T)).$$

This shows that

$$g_j(\bar{y}(T)) \leq \frac{V}{T}|w_j(T) - w_j(0)| \quad j \in \{1, \ldots, J\}. \quad (21)$$

Substituting (21) and (19) into (20) proves (16). ■

Theorem 1 can be interpreted when $\|\lambda(T)\| = \|(w(T), z(T))\|$ is bounded from above by some finite constant to mean that the deviation from optimality (15) is bounded from above by $O(V/T + 1/V)$, and the constraint violation (16) is bounded above by $O(V/T)$. To have both bounds be within $O(\epsilon)$, we set $V = 1/\epsilon$ and $T = 1/\epsilon^2$. Thus the convergence time of Algorithm 1 is $O(1/\epsilon^2)$. The next lemma shows that such a constant exists when the Slater condition holds.

Lemma 1: When $V \geq 1$, $w_j(0) = z_i(0) = 0$ for all i and j, then under Algorithm 1, the Slater condition implies there is a constant $D > 0$ (independent of V) such that

$$\|\lambda(t)\| = \sqrt{\sum_{j=1}^{J} w_j(t)^2 + \sum_{i=1}^{I} z_i(t)^2} \leq D \quad \text{for all } t.$$

Proof: From (14) and $V \geq 1$, if $\|\lambda(t)\| \geq (C + F - f^{(\min)})/\eta$ where $f^{(\min)} = \inf_{y\in\mathcal{Y}} f(y)$, then we have

$$\frac{V}{2}\Big[\|\lambda(t+1)\|^2 - \|\lambda(t)\|^2\Big] \leq \frac{C}{V} + F - f(y(t)) - \eta\|\lambda(t)\| \leq 0$$

This implies that:

$$\|\lambda(t)\| \leq (C + F - f^{(\min)})/\eta + \|\lambda(t+1) - \lambda(t)\|.$$

To complete the proof, note that $\|\lambda(t+1) - \lambda(t)\| \leq \sqrt{2C}/V$ from (10). Since $V \geq 1$, letting $D \triangleq (C + F - f^{(\min)})/\eta + \sqrt{2C}$ proves the lemma. ■

This section shows that Algorithm 1 generates a sequence of decisions that achieves an $O(\epsilon)$-optimal solution within $O(1/\epsilon^2)$ iterations. The next section shows that it is possible to generate an $O(\epsilon)$-optimal achieving sequence of decisions within $O(1/\epsilon)$ iterations and $O(1/\epsilon^{1.5})$ iterations (depending on a curvature property of the problem) by analyzing a *transient phase* and a *steady state phase* of Algorithm 1.

IV. CONVERGENCE OF TRANSIENT AND STEADY STATE PHASES

We analyze the convergence time in the case when the dual function satisfies a *locally-polyhedral* assumption and the case when it satisfies a *locally-smooth* assumption. Both cases use the following mild assumption:

Assumption 1: The dual formulation (7) has a unique Lagrange multiplier denoted by $\lambda^* \triangleq (w^*, z^*)$.

This assumption is assumed throughout Section IV, and replaces the Slater assumption (which is no longer needed). Note that this is a mild assumption when practical systems are considered, e.g., [15], [18].

Lemma 2: Let $\{\lambda(t)\}_{t=0}^{\infty}$ be a sequence generated by Algorithm 1. The following relation holds:

$$\|\lambda(t+1) - \lambda^*\|^2 \leq \|\lambda(t) - \lambda^*\|^2 + \frac{2}{V}[d(\lambda(t)) - d(\lambda^*)] + \frac{2C}{V^2}, \quad t \in \{0, 1, 2, \ldots\}. \quad (22)$$

Proof: Recall that $\lambda(t) = (w(t), z(t))$. Define $h(t) \triangleq (g(y(t)), x(t) - y(t))$ as the vector of the constraint functions. From the non-expansive property, we have that

$$\begin{aligned}
&\|\lambda(t+1) - \lambda^*\|^2 \\
&\leq \big\|\lambda(t) + \tfrac{1}{V}h(t) - \lambda^*\big\|^2 \\
&= \|\lambda(t) - \lambda^*\|^2 + \tfrac{1}{V^2}\|h(t)\|^2 + \tfrac{2}{V}[\lambda(t) - \lambda^*]^\top h(t) \\
&\leq \|\lambda(t) - \lambda^*\|^2 + \tfrac{2C}{V^2} + \tfrac{2}{V}[d(\lambda(t)) - d(\lambda^*)], \qquad (23)
\end{aligned}$$

where the last inequality uses the definition of C and the concavity of the dual function (6), i.e, $d(\lambda_1) \leq d(\lambda_2) + \partial d(\lambda_2)^\top[\lambda_1 - \lambda_2]$ for any $\lambda_1, \lambda_2 \in \Pi$, and $\partial d(\lambda(t)) = h(t)$. ■

A. Locally-Polyhedral Dual Function

Throughout Section IV-A, the dual function (6) is assumed to have a locally-polyhedral property, introduced in [15], as stated in Assumption 2. A dual function with this property is illustrated in Figure 1. The property holds when f and g_j for every j are either linear or piece-wise linear.

Assumption 2: There exists an $L_{\text{p}} > 0$ such that the dual function (6) satisfies

$$d(\lambda^*) \geq d(\lambda) + L_{\text{p}}\|\lambda - \lambda^*\| \quad \text{for all } \lambda \in \Pi \quad (24)$$

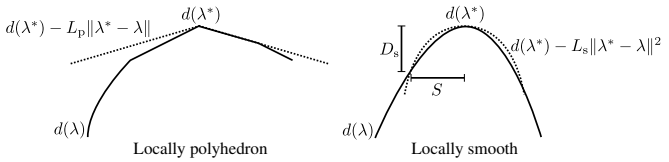

Fig. 1. Illustration of locally-polyhedral and locally-smooth functions

where λ^* is the unique Lagrange multiplier.

The "p" subscript in $L_{\rm p}$ represents "polyhedral." Furthermore, concavity of dual function (6) ensures that if this property holds locally about λ^*, it also holds globally for all $\lambda \in \Pi$ (see Figure 1).

The behavior of the generated dual variables with dual function satisfying the locally-polyhedral assumption can be described as follows. Define

$$B_{\rm p}(V) \triangleq \max\left\{\tfrac{L_{\rm p}}{2V}, \tfrac{2C}{VL_{\rm p}}\right\}.$$

Lemma 3: Under Assumptions 1 and 2, whenever $\|\lambda(t) - \lambda^*\| \geq B_{\rm p}(V)$, it follows that

$$\|\lambda(t+1) - \lambda^*\| - \|\lambda(t) - \lambda^*\| \leq -\frac{L_{\rm p}}{2V}. \tag{25}$$

Proof: Suppose the following condition holds

$$\frac{2}{V}[d(\lambda(t)) - d(\lambda^*)] + \frac{2C}{V^2} \leq -\frac{L_{\rm p}}{V}\|\lambda(t) - \lambda^*\| + \frac{L_{\rm p}^2}{4V^2}, \tag{26}$$

then the inequality (22) in Lemma 2 becomes

$$\begin{aligned}\|\lambda(t+1) - \lambda^*\|^2 &\leq \|\lambda(t) - \lambda^*\|^2 - \frac{L_{\rm p}}{V}\|\lambda(t) - \lambda^*\| + \frac{L_{\rm p}^2}{4V^2} \\ &= \left[\|\lambda(t) - \lambda^*\| - \frac{L_{\rm p}}{2V}\right]^2.\end{aligned}$$

It follows that if $\|\lambda(t) - \lambda^*\| \geq B_{\rm p}(V) \geq \frac{L_{\rm p}}{2V}$, then inequality (25) holds.

It requires to show that condition (26) holds when $\|\lambda(t) - \lambda^*\| \geq B_{\rm p}(V)$. Note that condition (26) holds when

$$d(\lambda(t)) - d(\lambda^*) \leq -\frac{C}{V} - \frac{L_{\rm p}}{2}\|\lambda(t) - \lambda^*\|.$$

By the locally-polyhedral property (24), if $-L_{\rm p}\|\lambda(t) - \lambda^*\| \leq -\frac{C}{V} - \frac{L_{\rm p}}{2}\|\lambda(t) - \lambda^*\|$, then the above inequality holds. This means that condition (26) holds when $\|\lambda(t) - \lambda^*\| \geq \frac{2C}{VL_{\rm p}}$. This proves the lemma. ■

Lemma 3 implies that, if the distance between $\lambda(t)$ and λ^* is at least $B_{\rm p}(V)$, the successor $\lambda(t+1)$ will be closer to λ^*. This suggests the existence of a convergence set in which a subsequence of $\{\lambda(t)\}_{t=0}^{\infty}$ resides. Note that $\sqrt{2C}/V$ bounds $\|\lambda(t+1) - \lambda(t)\|$ for all t as in (10).

The steady state of Algorithm 1 is defined from this set. This convergence set is defined as

$$\mathcal{R}_{\rm p}(V) = \left\{\lambda \in \Pi : \|\lambda - \lambda^*\| \leq B_{\rm p}(V) + \tfrac{\sqrt{2C}}{V}\right\}. \tag{27}$$

Let $T_{\rm p}$ be the first iteration that a generated dual variable enters this set:

$$T_{\rm p} = \operatorname*{arginf}_{t \geq 0}\{\lambda(t) \in \mathcal{R}_{\rm p}(V)\}. \tag{28}$$

Intuitively, $T_{\rm p}$ is the end of the transient phase and is the beginning of the steady state phase.

Lemma 4: Under Assumptions 1 and 2, $T_{\rm p} \leq O(V)$.

Proof: Since $\|\lambda(0) - \lambda^*\|$ is a constant, Lemma 3 proves the claim. ■

Then we show that dual variables generated after iteration $T_{\rm p}$ never leave $\mathcal{R}_{\rm p}(V)$.

Lemma 5: Under Assumptions 1 and 2, the generated dual variables from Algorithm 1 satisfy $\lambda(t) \in \mathcal{R}_{\rm p}(V)$ for all $t \geq T_{\rm p}$.

Proof: We prove the lemma by induction. First we note that $\lambda(T_{\rm p}) \in \mathcal{R}_{\rm p}(V)$ by the definition of $T_{\rm p}$. Suppose that $\lambda(t) \in \mathcal{R}_{\rm p}(V)$. Then two cases are considered.

i) If $\|\lambda(t) - \lambda^*\| \geq B_{\rm p}(V)$, it follows from (25) that

$$\|\lambda(t+1) - \lambda^*\| \leq \|\lambda(t) - \lambda^*\| - \frac{L_{\rm p}}{2V} \leq B_{\rm p}(V) + \frac{\sqrt{2C}}{V}.$$

ii) If $\|\lambda(t) - \lambda^*\| \leq B_{\rm p}(V)$, it follows from the triangle inequality that

$$\begin{aligned}\|\lambda(t+1) - \lambda^*\| &\leq \|\lambda(t+1) - \lambda(t)\| + \|\lambda(t) - \lambda^*\| \\ &\leq \frac{\sqrt{2C}}{V} + B_{\rm p}(V),\end{aligned}$$

by (10) and the assumption of $\|\lambda(t) - \lambda^*\|$. Hence, $\lambda(t+1) \in \mathcal{R}_{\rm p}(V)$ in both cases. This proves the lemma by induction. ■

Finally, a convergence result is ready to be stated. Let

$$\overline{a_{T_{\rm p}}}(T) = \frac{1}{T}\sum_{t=T_{\rm p}}^{T_{\rm p}+T-1} a(t)$$

be an average of sequence $\{a(t)\}_{t=T_{\rm p}}^{T_{\rm p}+T-1}$ that starts from $T_{\rm p}$.

Theorem 2: Under Assumptions 1 and 2, for $T > 0$, let $\{x(t), w(t)\}_{t=T_{\rm p}}^{\infty}$ be a subsequence generated by Algorithm 1, where $T_{\rm p}$ is defined in (28). The following bounds hold:

$$\begin{aligned}f(\overline{x_{T_{\rm p}}}(T)) - f^{(\rm opt)} &\leq \frac{C}{V} + \frac{2VM}{T}\left[\frac{\sqrt{2C}}{V} + B_{\rm p}(V)\right] \\ &+ \frac{V}{2T}\left\{\left[\frac{\sqrt{2C}}{V} + B_{\rm p}(V)\right]^2 + 4\|\lambda^*\|\left[\frac{\sqrt{2C}}{V} + B_{\rm p}(V)\right]\right\}\end{aligned} \tag{29}$$

$$g_j(\overline{x_{T_{\rm p}}}(T)) \leq \frac{2V(1+M)}{T}\left[\frac{\sqrt{2C}}{V} + B_{\rm p}(V)\right], \quad j \in \{1, \ldots, J\}. \tag{30}$$

Proof: The first part of the theorem follows from (15) with the average starting from $T_{\rm p}$ that

$$\begin{aligned}f(\overline{x_{T_{\rm p}}}(T)) - f^{(\rm opt)} &\leq \frac{C}{V} + \frac{V}{2T}\Big[\|\lambda(T_{\rm p})\|^2 - \|\lambda(T_{\rm p}+T)\|^2\Big] \\ &+ \frac{VM}{T}\|z(T_{\rm p}+T) - z(T_{\rm p})\|.\end{aligned} \tag{31}$$

For any $\lambda \in \Pi$, it holds that:

$$\|\lambda\|^2 = \|\lambda - \lambda^*\|^2 + \|\lambda^*\|^2 + 2[\lambda - \lambda^*]^\top \lambda^*.$$

The second term on the right-hand-side of (31) can be upper bounded by applying this equality.

$$\begin{aligned}
&\|\lambda(T_{\rm p})\|^2 - \|\lambda(T_{\rm p}+T)\|^2 \\
&\quad = \|\lambda(T_{\rm p})-\lambda^*\|^2 + 2[\lambda(T_{\rm p})-\lambda^*]^\top\lambda^* \\
&\qquad - \|\lambda(T_{\rm p}+T)-\lambda^*\|^2 - 2[\lambda(T_{\rm p}+T)-\lambda^*]^\top\lambda^* \\
&\quad \le \|\lambda(T_{\rm p})-\lambda^*\|^2 + 2[\lambda(T_{\rm p})-\lambda(T_{\rm p}+T)]^\top\lambda^* \\
&\quad \le \|\lambda(T_{\rm p})-\lambda^*\|^2 + 2\|\lambda(T_{\rm p})-\lambda(T_{\rm p}+T)\|\|\lambda^*\| \qquad (32)
\end{aligned}$$

From Lemma 5, the first term of (32) is bounded by $\|\lambda(T_{\rm p})-\lambda^*\|^2 \le [\sqrt{2C}/V + B_{\rm p}(V)]^2$. From triangle inequality and Lemma 5, the last term of (32) is bounded by

$$\begin{aligned}
\|\lambda(T_{\rm p}+T)-\lambda(T_{\rm p})\| &\le \|\lambda(T_{\rm p}+T)-\lambda^*\| + \|\lambda^*-\lambda(T_{\rm p})\| \\
&\le 2\left[\sqrt{2C}/V + B_{\rm p}(V)\right]. \qquad (33)
\end{aligned}$$

Therefore, inequality (32) is bounded from above by $[\sqrt{2C}/V + B_{\rm p}(V)]^2 + 4\|\lambda^*\|[\sqrt{2C}/V + B_{\rm p}(V)]$. Substituting this bound into (31) and using the fact that $\|z(T_{\rm p}+T)-z(T_{\rm p})\| \le \|\lambda(T_{\rm p}+T)-\lambda(T_{\rm p})\| \le 2[\sqrt{2C}/V + B_{\rm p}(V)]$ proves the first part of the theorem.

The last part follows from (16) that

$$\begin{aligned}
g_j(\overline{x_{T_{\rm p}}}(T)) &\le \frac{V}{T}|w_j(T_{\rm p}+T)-w_j(T_{\rm p})| \\
&\quad + \frac{VM}{T}\|z(T_{\rm p}+T)-z(T_{\rm p})\|.
\end{aligned}$$

Since $|w_j(T_{\rm p}+T)-w_j(T_{\rm p})|$ and $\|z(T_{\rm p}+T)-z(T_{\rm p})\|$ are bounded above by $\|\lambda(T_{\rm p}+T)-\lambda(T_{\rm p})\|$, the above inequality is upper bounded by

$$\begin{aligned}
g_j(\overline{x_{T_{\rm p}}}(T)) &\le \frac{V(1+M)}{T}\|\lambda(T_{\rm p}+T)-\lambda(T_{\rm p})\| \\
&\le \frac{2V(1+M)}{T}\left[\frac{\sqrt{2C}}{V} + B_{\rm p}(V)\right],
\end{aligned}$$

where the last inequality uses relation (33). This proves the last part of the theorem. ■

Theorem 2 can be interpreted as follows. The deviation from the optimality value (29) is bounded above by $O(1/V+1/T)$. The constraint violation (30) is bounded above by $O(1/T)$. To have both bounds be within $O(\epsilon)$, we set $V = 1/\epsilon$ and $T = 1/\epsilon$, and the convergence time of Algorithm 1 is $O(1/\epsilon)$. Note that both bounds consider the average starting after reaching the steady state at time $T_{\rm p}$, and this transient time $T_{\rm p}$ is at most $O(1/\epsilon)$.

B. Locally-Smooth Dual Function

Throughout Section IV-B, the dual function (6) is assumed to have a locally-smooth property, introduced in [15], as stated in Assumption 3 and illustrated in Figure 1.

Assumption 3: Let λ^* be the unique Largrange multiplier, there exist $S > 0$ and $L_{\rm s} > 0$ such that whenever $\lambda \in \Pi$ and $\|\lambda-\lambda^*\| \le S$, dual function (6) satisfies

$$d(\lambda^*) \ge d(\lambda) + L_{\rm s}\|\lambda-\lambda^*\|^2. \qquad (34)$$

Also, there exists $D_{\rm s} > 0$ such that whenever $\lambda \in \Pi$ and $d(\lambda^*) - d(\lambda) \le D_{\rm s}$, dual variable satisfies $\|\lambda-\lambda^*\| \le S$.

TABLE I
CONVERGENCE TIMES

	General	Polyhedron	Smooth
Transient state	0	$O(1/\epsilon)$	$O(1/\epsilon^{1.5})$
Steady state	$O(1/\epsilon^2)$	$O(1/\epsilon)$	$O(1/\epsilon^{1.5})$

The "s" subscript in $L_{\rm s}$ represents "smooth."

Using a similar proof process as in Section IV-A, the convergence result under the locally-smooth property is as follows. Define the smooth counterparts of $B_{\rm p}(V)$ and $T_{\rm p}(V)$:

$$\begin{aligned}
B_{\rm s}(V) &\triangleq \max\left\{\tfrac{1}{V^{1.5}}, \tfrac{\sqrt{V}+\sqrt{V+4L_{\rm s}CV}}{2L_{\rm s}V}\right\} \\
\mathcal{R}_{\rm s}(V) &= \left\{\lambda\in\Pi : \|\lambda-\lambda^*\| \le B_{\rm s}(V) + \tfrac{\sqrt{2C}}{V}\right\} \qquad (35) \\
T_{\rm s} &= \operatorname*{arginf}_{t\ge 0}\{\lambda(t)\in\mathcal{R}_{\rm s}(V)\}.
\end{aligned}$$

Theorem 3: Under Assumptions 1 and 3, when V is sufficiently large and $B_{\rm s}(V) + \frac{\sqrt{2C}}{V} < S$, for $T > 0$, let $\{x(t), w(t)\}_{t=T_{\rm s}}^{\infty}$ be a subsequence generated by Algorithm 1, where $T_{\rm s}$ is defined in (35). The following bounds hold:

$$\begin{aligned}
f(\overline{x_{T_{\rm s}}}(T)) - f^{(\rm opt)} &\le \frac{C}{V} + \frac{2VM}{T}\left[\frac{\sqrt{2C}}{V} + B_{\rm s}(V)\right] \\
&\quad + \frac{V}{2T}\left\{\left[\frac{\sqrt{2C}}{V} + B_{\rm s}(V)\right]^2 + 2\|\lambda^*\|\left[\frac{\sqrt{2C}}{V} + B_{\rm s}(V)\right]\right\} \qquad (36) \\
g_j(\overline{x_{T_{\rm s}}}(T)) &\le \frac{2V(1+M)}{T}\left[\frac{\sqrt{2C}}{V} + B_{\rm s}(V)\right], \\
&\qquad j\in\{1,\dots,J\}. \qquad (37)
\end{aligned}$$

Proof: Please see the full proof in [19]. ■

Theorem 3 can be interpreted as follows. The deviation from the optimality (36) is bounded above by $O(1/V + \sqrt{V}/T)$. The constraint violation (37) is bounded above by $O(\sqrt{V}/T)$. To have both bounds be within $O(\epsilon)$, we set $V = 1/\epsilon$ and $T = 1/\epsilon^{1.5}$, and the convergence time of Algorithm 1 is $O(1/\epsilon^{1.5})$. Note that both bounds consider the average starting after reaching the steady state at time $T_{\rm s}$, and this transient time $T_{\rm s}$ is at most $O(1/\epsilon^{1.5})$, which has been shown in [19].

C. Summary of Convergence Results

The results in Theorems 1, 2, and 3 (denoted by General, Polyhedron, and Smooth) are summarized in Table I. Note that the general convergence time is considered to be in the steady state from the beginning.

D. Staggered Time Averages

In order to take advantage of the improved convergence rates, computing time averages must be started after the transient phase. To achieve this performance without determining the exact end time of the transient phase, time averages can be restarted over successive frames whose frame lengths increase geometrically. For example, if one triggers a restart at times 2^k for integers k, then a restart is guaranteed to occur within a factor of 2 of the time of the actual end of the transient phase.

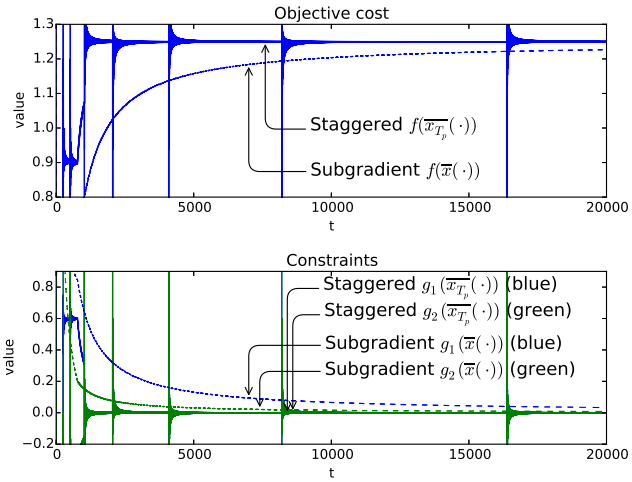

Fig. 2. Iterations solving problem (38) with $f(x) = 1.5x_1 + x_2$

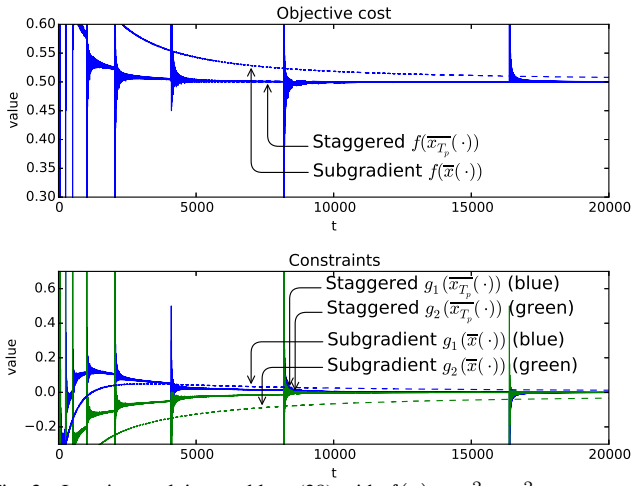

Fig. 3. Iterations solving problem (38) with $f(x) = x_1^2 + x_2^2$

V. Sample Problems

This section illustrates the convergence times of the time-average Algorithm 1 under locally-polyhedral and locally-smooth assumptions. A considered formulation is

$$\begin{aligned} \text{Minimize} \quad & f(\bar{x}) \\ \text{Subject to} \quad & 2\bar{x}_1 + \bar{x}_2 \geq 1.5, \quad \bar{x}_1 + 2\bar{x}_2 \geq 1.5 \\ & x_1(t), x_2(t) \in \{0, 1, 2, 3\}, \quad t \in \{0, 1, 2, \dots\} \end{aligned} \tag{38}$$

where function f will be given for different cases.

Under the locally-polyhedral assumption, let $f(x) = 1.5x_1 + x_2$ be the objective function of problem (38). In this setting, the optimal value is 1.25 when $\bar{x}_1 = \bar{x}_2 = 0.5$. Figure 2 shows the values of objective and constraint functions of time-averaged solutions. It is easy to see the faster convergence time $O(1/\epsilon)$ from the polyhedral result ($T_\text{p} = 2048$) compared to a general result with convergence time $O(1/\epsilon^2)$.

Under the locally-smooth assumption, let $f(x) = x_1^2 + x_2^2$ be the objective function of problem (38). Note that the optimal value of this problem is 0.5 where $\bar{x}_1 = \bar{x}_2 = 0.5$. Figure 3 shows the values of objective and constraint functions of time-averaged solutions. The smooth result starts the average from $(T_\text{s} =)8192^\text{th}$ iterations. It is easy to see that the general result converges slower than the smooth result. This illustrates the difference between $O(1/\epsilon^2)$ and $O(1/\epsilon^{1.5})$.

VI. Conclusion

We consider the time-average optimization problem with a non-convex (possibly discrete) decision set. We show that the problem has a corresponding (one-shot) convex optimization formulation. This connects the Lyapunov optimization technique and convex optimization theory. Using convex analysis we prove a general convergence time of $O(1/\epsilon^2)$ when the Slater condition holds. Under an assumption on the uniqueness of a Lagrange multiplier, we prove that faster convergence times $O(1/\epsilon)$ and $O(1/\epsilon^{1.5})$ are possible for locally-polyhedral and locally-smooth problems.

References

[1] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, Jan. 2007.

[2] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *Automatic Control, IEEE Transactions on*, vol. 54, no. 1, pp. 48–61, Jan 2009.

[3] ——, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM Journal on Optimization*, vol. 19, no. 4, 2009.

[4] A. Nemirovskii and D. Yudin, "Cesaro convergence of the gradient method for approximation of saddle points of convex-concave functions," *Doklady AN SSSR 239*, 1978.

[5] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Mathematical Programming*, vol. 120, no. 1, 2009.

[6] M. Neely, "Distributed and secure computation of convex programs over a network of connected processors," *DCDIS Conf.*, Jul. 2005.

[7] M. Zhu and S. Martinez, "An approximate dual subgradient algorithm for multi-agent non-convex optimization," *Automatic Control, IEEE Transactions on*, vol. 58, no. 6, pp. 1534–1539, June 2013.

[8] M. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, 2010.

[9] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Springer Netherlands, 2004.

[10] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," *submitted to SIAM Journal on Optimization*, 2008.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, USA: Cambridge University Press, 2004.

[12] A. Beck, A. Nedić, A. Ozdaglar, and M. Teboulle, "An o(1/k) gradient method for network resource allocation problems," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 64–73, Mar. 2014.

[13] E. Wei and A. Ozdaglar, "On the o(1/k) convergence of asynchronous distributed alternating direction method of multipliers," in *2013 IEEE Global Conference on Signal and Information Processing*, Dec. 2013.

[14] J. Liu, C. Xia, N. Shroff, and H. Sherali, "Distributed cross-layer optimization in wireless networks: A second-order approach," in *INFOCOM, 2013 Proceedings IEEE*, Apr 2013.

[15] L. Huang and M. Neely, "Delay reduction via lagrange multipliers in stochastic network optimization," *Automatic Control, IEEE Transactions on*, vol. 56, no. 4, Apr. 2011.

[16] D. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization*. Athena Scientific, 2003.

[17] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time varying wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, Jan. 2005.

[18] A. Eryilmaz and R. Srikant, "Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control," *Networking, IEEE/ACM Transactions on*, vol. 15, no. 6, Dec. 2007.

[19] S. Supittayapornpong, L. Huang, and M. J. Neely, "Time-average optimization with non-convex decision set and its convergence," *arXiv:1610.02617 [math.OC]*, Oct. 2016.