Time-Average Optimization with Nonconvex 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 nonconvex) 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 and operations research, where decisions can be constrained to discrete sets and time averages can represent bit rates, power expenditures, and so on. These problems can be solved by Lyapunov optimization. This paper shows that 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. However, when the problem has a unique vector of Lagrange multipliers, the algorithm is shown to have a transient phase and a steady state phase. By restarting the time averages after the transient phase, the total convergence time is improved to $O(1/\epsilon)$ under a locally-polyhedron assumption, and to $O(1/\epsilon^{1.5})$ under a locally-smooth assumption.**

I. INTRODUCTION

Convex optimization is often used to optimally control communication networks and distributed multi-agent systems (see [1] and references therein). This framework utilizes both convexity properties of an objective function and a feasible decision set. However, various systems have inherent discrete (and hence nonconvex) decision sets. For example, a packet switch system makes a binary (0/1) decision about connecting a given link. Further, a wireless system might constrain transmission rates to a finite set corresponding to a fixed set of coding options. This discreteness restrains the application of convex optimization.

Let I and J be positive integers. This paper considers *time-average optimization* where decision vectors $x(t) = (x_1(t), \dots, x_I(t))$ are chosen sequentially over time slots $t \in \{0, 1, 2, \dots\}$ to solve the following problem:

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

Here $\mathcal{X}$ is a closed and bounded subset of $\mathbb{R}^I$ (possibly nonconvex and discrete), $\overline{\mathcal{X}}$ is its convex hull, $f : \overline{\mathcal{X}} \to \mathbb{R}$ and $g_j : \overline{\mathcal{X}} \to \mathbb{R}$ are convex functions, and $\bar{x} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} x(t)$ is an average of decisions made.

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.

S. Supittayapornpong and M. J. Neely are with Electrical Engineering Department, University of Southern California, 3740 McClintock Ave., Los Angeles, CA 90089-2565, supittay@usc.edu, mjneely@usc.edu

L. Huang is with Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China, 100084, longbohuang@tsinghua.edu.cn

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

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

However, an optimal solution to formulation (2) may not be in the nonconvex decision set $\mathcal{X}$. Nevertheless, problems (1) and (2) have the same optimal value.

Although there have been several techniques utilizing time-averaged solutions [2], [3], [4], those works are limited to convex formulations. This paper is inspired by the Lyapunov optimization technique [5] which solves stochastic and time-averaged optimization problems. This paper removes the stochastic characteristic and focuses on the connection between the technique and a general convex optimization. This allows a convergence time analysis of a *drift-plus-penalty* algorithm that solves problem (1). Further, this paper shows that faster convergence can be achieved by starting time averages after a suitable transient 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, the optimal first-order method has $O(1/\sqrt{\epsilon})$ convergence time [6], [7], while gradient (without strong convexity of objective function) and subgradient methods take $O(1/\epsilon)$ and $O(1/\epsilon^2)$ respectively [8], [3], [4]. Two $O(1/\epsilon)$ first-order methods for constrained optimization are developed in [9], [10], but the results rely on special convex formulations. A second-order method for constrained optimization [11] has a fast convergence rate but rely on special a convex formulation. All of these results rely on convexity assumption 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 for general problems. However, inspired by results in [12], 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 restarting the time averages after the transient phase. Specifically, when a dual function satisfies a *locally-polyhedron* 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. Furthermore, simulations show that these fast convergence times are robust even without the uniqueness assumption. We conjecture that our results hold even when the uniqueness assumption is removed.

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 result 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, including cases when the uniqueness condition fails. In all cases, the method of restarting time averages later than time 0 is shown to significantly improve convergence.

II. TIME-AVERAGE OPTIMIZATION

In order to solve problem (1), an auxiliary problem with a similar solution is formulated. Then Lyapunov optimization [5] is applied on this auxiliary problem.

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 $j \in \{1,\dots,J\}$ extend as real-valued 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). Further, choosing $\mathcal{Y}$ to have a simple structure helps to simplify 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 that there is a constant $M > 0$ such that for all $x, y \in \mathcal{Y}$:

$$|f(x) - f(y)| \le M\|x - y\| \tag{3}$$
$$|g_j(x) - g_j(y)| \le M\|x - y\| \tag{4}$$

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 $j \in \{1,\dots,J\}$, and is such that $\hat{x}$ is in the interior of set $\mathcal{Y}$. This is a *Slater condition* that, among other things, ensures the constraints are feasible for the problem of interest.

C. Auxiliary Problem

For functions $a(x(t))$ of a vector $x(t)$, let notation $\overline{a(x)} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} a(x(t))$ represent an average of function values. An *auxiliary formulation* of problem (1) is

$$\begin{array}{lll} \text{Minimize} & \overline{f(y)} & \\ \text{Subject to} & \overline{g_j(y)} \le 0 & j \in \{1,\dots,J\} \\ & \overline{x_i} = \overline{y_i} & i \in \{1,\dots,I\} \\ & x(t) \in \mathcal{X}, \quad y(t) \in \mathcal{Y} & t \in \{0,1,2,\dots\}. \end{array} \tag{5}$$

The auxiliary vector $y(t)$ is introduced so it can be chosen in the convexified set $\overline{\mathcal{X}} \subseteq \mathcal{Y}$. Constraint $\overline{x}_i = \overline{y}_i$ ensures that vectors $x(t)$ and $y(t)$ have the same time averages. For ease of notation, let $g(y) = (g_1(y),\dots,g_J(y))$ denote a J-column vector of functions $g_j(y)$.

Recall that problems (1) and (2) share the same optimal objective cost. Let $f^{(\text{opt})}$ denote that optimal cost. The following theorem is proven via Jensen's inequality in [5]:

Theorem 1: The optimal objective function value in problem (5) is also $f^{(\text{opt})}$. Further, if $\{x^*(t), y^*(t)\}_{t=0}^{\infty}$ is an optimal solution to problem (5), then $\{x^*(t)\}_{t=0}^{\infty}$ is an optimal solution to problem (1).

D. Lyapunov optimization

Problem (5) can be solved by the Lyapunov optimization technique [5]. Define $W_j(t)$ and $Z_i(t)$ as *virtual queues* of constraints $\overline{g_j(y)} \le 0$ and $\overline{x_i} = \overline{y_i}$, with the update equations:

$$W_j(t+1) = [W_j(t) + g_j(y(t))]_+ \qquad j \in \{1,\dots,J\} \tag{6}$$
$$Z_i(t+1) = Z_i(t) + x_i(t) - y_i(t) \qquad i \in \{1,\dots,I\}, \tag{7}$$

where the operator $[\cdot]_+$ is a projection to a corresponding non-negative orthant. For ease of notation, let $W(t) \triangleq (W_1(t),\dots,W_J(t))$ and $Z(t) \triangleq (Z_1(t),\dots,Z_I(t))$ be vectors of $W_j(t)$'s and $Z_i(t)$'s respectively, and let notation $A^\top$ denote the transpose of vector A.

Define a Lyapunov function as:

$$L(t) \triangleq \frac{1}{2}\|W(t)\|^2 + \frac{1}{2}\|Z(t)\|^2$$

Define the *Lyapunov drift* as $\Delta(t) \triangleq L(t+1) - L(t)$.

Lemma 1: For every t, the Lyapunov drift is bounded by

$$\Delta(t) \le C_3 + W(t)^\top g(y(t)) + Z(t)^\top [x(t) - y(t)],$$

where $C_3 = (C_1^2 + C_2^2)/2$ and $C_1 \triangleq \sup_{y\in\mathcal{Y}} \|g(y)\|$ and $C_2 \triangleq \sup_{x\in\mathcal{X}, y\in\mathcal{Y}} \|x - y\|$ are bounded values, as $\mathcal{X}$ and $\mathcal{Y}$ are bounded.

Proof: Equation (6) gives $\|W(t+1)\|^2 \le \|W_j(t) + g_j(y(t))\|^2$ and $\|W(t+1)\|^2 - \|W(t)\|^2 \le 2W(t)^\top g(y(t)) + \|g(y(t))\|^2$. Similarly, equation (7) gives $\|Z(t+1)\|^2 - \|Z(t)\|^2 = 2Z(t)^\top[x(t) - y(t)] + \|x(t) - y(t)\|^2$. Summing the last two relations and using the definitions of C_1 and C_2 yield $2\Delta(t) \le 2W(t)^\top g(y(t)) + 2Z(t)^\top[x(t) - y(t)] + C_1^2 + C_2^2$, which proves the lemma. ∎

Let $V > 0$ be a real number (used as a parameter in the Lyapunov optimization technique). The drift-plus-penalty expression is defined by $\Delta(t) + Vf(y(t))$. Applying Lemma 1, the drift-plus-penalty expression is bounded, for all t, by

$$\Delta(t) + Vf(y(t)) \le C_3 + W(t)^\top g(y(t)) + Z(t)^\top[x(t) - y(t)] + Vf(y(t)). \tag{8}$$

E. Drift-plus-penalty algorithm

A Lyapunov optimization algorithm, at every iteration, minimizes the right-hand-side of inequality (8) with respect to $x(t) \in \mathcal{X}$ and $y(t) \in \mathcal{Y}$ and updates the virtual queues $W(t)$ and $Z(t)$ with equations (6) and (7). Let W_0 and Z_0 be the initialized values of $W(0)$ and $Z(0)$. Then, the algorithm is summarized in Algorithm 1.

```
Initialize W(0) = W_0, Z(0) = Z_0.
for t = 0, 1, 2, ... do
    x(t) = argmin_{x∈X} Z(t)^⊤ x
    y(t) = argmin_{y∈Y} [V f(y) + W(t)^⊤ g(y) − Z(t)^⊤ y]
    W(t+1) = [W(t) + g(y(t))]_+
    Z(t+1) = Z(t) + x(t) − y(t)
end
```

Algorithm 1: Drift-plus-penalty algorithm solving (5).

Algorithm 1 generates sequence $\{x(t), y(t)\}_{t=0}^{\infty}$, which is an $O(\epsilon)$-optimal solution to the auxiliary problem (5) by setting $V = 1/\epsilon$ [5]. For an $O(\epsilon)$-optimal solution to the time-average problem (1), decision $x(t)$ made by Algorithm 1 is implemented every iteration t, which coincides with Theorem 1.

F. Relation to dual subgradient algorithm

It is interesting to note that the drift-plus-penalty algorithm is identical to a classic dual subgradient method [13] with a fixed stepsize $1/V$, with the exception that it takes a time average of $x(t)$ values. This was noted in [14], [12] for related problems. To see this for the problem of this paper, consider the following convex program, called the *embedded formulation* of the time-average problem (5):

$$\begin{aligned} \text{Minimize} \quad & f(y) && && (9)\\ \text{Subject to} \quad & g_j(y) \le 0 && j \in \{1, \dots, J\}\\ & x_i = y_i && i \in \{1, \dots, I\}\\ & x \in \overline{\mathcal{X}}, \quad y \in \mathcal{Y}. \end{aligned}$$

This problem 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)–(2), (5).

Now consider the dual of embedded formulation (9). Let vectors w and z be dual variables of the first and second constraints in problem (9), where the feasible set of (w, z) is denoted by $\Pi = \mathbb{R}_+^J \times \mathbb{R}^I$. A Lagrangian has the following expression:

$$\Lambda(x, y, w, z) = f(y) + w^\top g(y) + z^\top (x - y).$$

Define:

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

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

$$\begin{aligned} d(w, z) &= \inf_{x \in \overline{\mathcal{X}}, y \in \mathcal{Y}} \Lambda(x, y, w, z) \qquad (10)\\ &= f(y^*(w, z)) + w^\top g(y^*(w, z)) + z^\top [x^*(z) - y^*(w, z)], \end{aligned}$$

and its subgradient is [13]:

$$\frac{\partial d}{\partial w}(w, z) = g(y^*(w, z)), \quad \frac{\partial d}{\partial z}(w, z) = x^*(z) - y^*(w, z)$$

Finally, the dual formulation of embedded problem (9) is

$$\begin{aligned} \text{Maximize} \quad & d(w, z) && (11)\\ \text{Subject to} \quad & (w, z) \in \Pi. \end{aligned}$$

Let the optimal value of problem (11) be d^*. Since problem (9) is convex, the duality gap is zero, and $d^* = f^{(\text{opt})}$. Problem (11) can be treated by a dual subgradient method [13] with a fixed stepsize $1/V$. This leads to Algorithm 2 summarized in the figure below, called the *dual subgradient algorithm*.

```
Initialize w(0) = W_0/V, z(0) = Z_0/V.
for t = 0, 1, 2, ... do
    x(t) = arg inf_{x∈X̄} z(t)^⊤ x    (with x(t) ∈ X)
    y(t) = arg inf_{y∈Y} [f(y) + w(t)^⊤ g(y) − z(t)^⊤ y]
    w(t+1) = [w(t) + (1/V) g(y(t))]_+
    z(t+1) = z(t) + (1/V)[x(t) − y(t)]
end
```

Algorithm 2: Dual subgradient algorithm solving (11).

Fix initial conditions $w(0) = W(0)/V$, $z(0) = Z(0)/V$. Then for all slots $\tau \in \{0, 1, 2, \dots\}$, Algorithms 1 and 2 always choose the same primal variables $x(\tau)$, $y(\tau)$, and their dual (virtual queue) variables are related by:

$$w(\tau) = W(\tau)/V, \quad z(\tau) = Z(\tau)/V \qquad (12)$$

To see this, assume (12) holds for all $\tau \in \{0, 1, \dots, t\}$ for some time $t \ge 0$. Then:

$$\arg\inf_{x \in \overline{\mathcal{X}}} z(t)^\top x = \arg\inf_{x \in \mathcal{X}} z(t)^\top x = \arg\inf_{x \in \mathcal{X}} Z(t)^\top x,$$

and

$$\begin{aligned} &\arg\inf_{y \in \mathcal{Y}} [f(y) + w(t)^\top g(y) - z(t)^\top y]\\ &= \arg\inf_{y \in \mathcal{Y}} [f(y) + \frac{1}{V} W(t)^\top g(y) - \frac{1}{V} Z(t)^\top y]\\ &= \arg\inf_{y \in \mathcal{Y}} [V f(y) + W(t)^\top g(y) - Z(t)^\top y]. \end{aligned}$$

So, the vectors $x(t), y(t)$ are the same on slot t under both algorithms. Then, it is easy to see the update equations for $W(t+1), Z(t+1)$ in Algorithm 1 and $w(t+1), z(t+1)$ in Algorithm 2 preserve the relationship $w(t+1) = W(t+1)/V$ and $z(t+1) = Z(t+1)/V$.

Traditionally, the dual subgradient algorithm of [13] is intended to produce primal vector estimates that converge to a desired result. However, this requires additional assumptions. Indeed, for our problem, 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, drift-plus-penalty theory of Lyapunov optimization can be used to ensure that the *time averages* of $x(t)$ and $y(t)$ converge as desired. Observing the relationship between Algorithms 1 and 2 enables one to integrate both duality and time-averaging concepts.

For the remainder of this paper, we use the notation $w(t)$ and $z(t)$ from Algorithm 2, with the update rule for $w(t+1)$ and $z(t+1)$ given there. For ease of notation, define $\lambda(t) \triangleq (w(t), z(t))$ as a concatenation of these vectors, and define $h(t)\triangleq(g(y(t)), x(t)-y(t))$ as the concatenation vector of the constraint functions.

Using this with $w(t) = W(t)/V$, $z(t) = Z(t)/V$ gives:

$$L(t) = \frac{V^2}{2}\|\lambda(t)\|^2.$$

Dividing the drift-plus-penalty inequality (8) by V and using this change of variables yields:

$$\begin{aligned}
&\frac{V}{2}\left[\|\lambda(t+1)\|^2 - \|\lambda(t)\|^2\right] + f(y(t)) \\
&\leq \frac{C_3}{V} + w(t)^\top g(y(t)) + z(t)^\top[x(t)-y(t)] + f(y(t)) \\
&= \frac{C_3}{V} + d(\lambda(t)) \qquad (13)
\end{aligned}$$

where the final inequality uses the definition of the dual function $d(\lambda(t)) = d(w(t), z(t))$ and the fact that Algorithm 2 chooses $x(t)$, $y(t)$ to minimize the expression

$$w(t)^\top g(y(t)) + z(t)^\top[x(t)-y(t)] + f(y(t)). \qquad (14)$$

G. Properties of the dual function

Because the dual function $d(\lambda(t))$ is the minimum of the expression (14) over $x(t) \in \overline{\mathcal{X}}$ and $y(t) \in \mathcal{Y}$, it satisfies:

$$d(\lambda(t)) \leq w(t)^\top g(y) + z(t)^\top[x-y] + f(y) \qquad (15)$$

for all $(x, y) \in \overline{\mathcal{X}} \times \mathcal{Y}$. Thus, the dual function $d(\lambda)$ has the following basic properties:[1]

- $d(\lambda) \leq f^{(\text{opt})}$ for all $\lambda \in \Pi$.
- If the Slater condition holds, then there are real numbers $F > 0$, $\eta > 0$ such that:
$$d(\lambda) \leq F - \eta\|\lambda\| \quad \text{for all } \lambda \in \Pi.$$
- If the Slater condition holds, then there is an optimal value $\lambda^* \in \Pi$, called a *Lagrange multiplier vector* [13], that maximizes $d(\lambda)$. Specifically, $d(\lambda^*) = f^{(\text{opt})}$.

The first two properties can be substituted into the modified drift-plus-penalty inequality (13) to ensure that, under Algorithm 2, the following inequalities hold for all time slots $t \in \{0, 1, 2, \dots\}$:

$$\frac{V}{2}\left[\|\lambda(t+1)\|^2 - \|\lambda(t)\|^2\right] + f(y(t)) \leq \frac{C_3}{V} + f^{(\text{opt})} \qquad (16)$$

$$\frac{V}{2}\left[\|\lambda(t+1)\|^2 - \|\lambda(t)\|^2\right] + f(y(t)) \leq \frac{C_3}{V} + F - \eta\|\lambda(t)\| \qquad (17)$$

[1]The first property follows by substituting the optimal solution $(x^{(\text{opt})}, x^{(\text{opt})})$ into the right-hand-side of (15), where $x^{(\text{opt})} \in \overline{\mathcal{X}}$ is a solution to problem (9). The second property can be shown by substituting $(\hat{x}+\hat{\boldsymbol{e}}, \hat{x})$ into the right-hand-side of (15), where the i$^{\text{th}}$-component of $\hat{\boldsymbol{e}}$, $\hat{\boldsymbol{e}}_i$ is a small negative value when $z_i(t)$ is positive; otherwise, $\boldsymbol{e}_i$ is a small positive value. The third property is standard for Lagrange multiplier theory.

III. GENERAL CONVERGENCE RESULT

Three useful lemmas are proved before the main theorem in this section. Define the average of variables $\{a(t)\}_{t=0}^{T-1}$ as

$$\bar{a}(T)\triangleq\frac{1}{T}\sum_{t=0}^{T-1} a(t).$$

Lemma 2: Let $\{x(t), y(t), w(t), z(t)\}_{t=0}^{\infty}$ be a sequence generated by Algorithm 2. For $T > 0$, we have

$$g_j(\bar{y}(T)) \leq \frac{V}{T}|w_j(T) - w_j(0)| \quad j \in \{1, \dots, J\} \qquad (18)$$

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

Proof: For the first part, the update equation of $w(t)$ in Algorithm 2 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 $t = 0, \dots, T-1$, we have $w_j(T) - w_j(0) \geq \frac{1}{V}\sum_{t=0}^{T-1} g_j(y(t))$. Dividing by T and using Jensen's inequality on the convexity of $g_j(\cdot)$ gives

$$g_j(\bar{y}(T)) \leq \frac{V}{T}[w_j(T) - w_j(0)] \leq \frac{V}{T}|w_j(T) - w_j(0)|,$$

which proves the upper bound (18).

The proof of equality (19) is similar and is omitted for brevity. ∎

Lemma 3: Let $\{x(t), y(t), w(t), z(t)\}_{t=0}^{\infty}$ be a sequence generated by Algorithm 2. For $T > 0$, it follows that

$$f(\bar{y}(T)) - f^{(\text{opt})} \leq \frac{V}{2T}\left[\|\lambda(0)\|^2 - \|\lambda(T)\|^2\right] + \frac{C_3}{V}. \qquad (20)$$

Proof: Relation (16) can be rewritten as

$$f(y(t)) - f^{(\text{opt})} \leq \frac{C_3}{V} + \frac{V}{2}\left[\|\lambda(t)\|^2 - \|\lambda(t+1)\|^2\right].$$

Summing from $t = 0, \dots, T-1$ and dividing by T give:

$$\frac{1}{T}\sum_{t=0}^{T-1} f(y(t)) - f^{(\text{opt})} \leq \frac{C_3}{V} + \frac{V}{2}\left[\|\lambda(0)\|^2 - \|\lambda(T)\|^2\right].$$

Using Jensen's inequality and the convexity of $f(\cdot)$ prove the lemma. ∎

Lemma 4: When $V \geq 1$, $w_j(0) = z_i(0) = 0$ for all i and j, then under Algorithm 2, the Slater condition implies there is a constant $D > 0$ (independent of V) such that $w_j(t) \leq D$ and $z_i(t) \leq D$ for all t and for all $i \in \{1, \dots, I\}$, $j \in \{1, \dots, J\}$.

Proof: From (17) and $V \geq 1$, if $\|\lambda(t)\| \geq (C_3 + F - f^{(\min)})/\eta$ where $f^{(\min)} = \inf_{y\in\mathcal{Y}} f(y)$, then we have

$$\begin{aligned}
\frac{V}{2}\left[\|\lambda(t+1)\|^2 - \|\lambda(t)\|^2\right] &\leq \frac{C_3}{V} + F - f(y(t)) - \eta\|\lambda(t)\| \\
&\leq 0.
\end{aligned}$$

But $\|\lambda(t+1) - \lambda(t)\|^2 = \|w(t+1) - w(t)\|^2 + \|z(t+1) - z(t)\|^2 \leq 2C_3/V^2$, and $\|\lambda(t+1) - \lambda(t)\| \leq \sqrt{2C_3}/V$ for all t. This implies that $\|\lambda(t)\| \leq (C_3 + F - f^{(\min)})/\eta + \sqrt{2C_3}/V$ for all t. Since $V \geq 1$, let $D\triangleq(C_3 + F - f^{(\min)})/\eta + \sqrt{2C_3}$ proves the lemma. ∎

Lemmas 2 and 3 provide bounds on $\bar{y}(T)$. The next result translates these to bounds on $\bar{x}(T)$.

Theorem 2: Let $\{x(t), w(t), z(t)\}_{t=0}^{\infty}$ be a sequence generated by Algorithm 2. For $T > 0$, we have

$$f(\bar{x}(T)) - f^{(\mathrm{opt})} \leq \frac{V}{2T}\left[\|\lambda(0)\|^2 - \|\lambda(T)\|^2\right] + \frac{C_3}{V} + \frac{VM}{T}\|z(T) - z(0)\| \quad (21)$$

$$g_j(\bar{x}(T)) \leq \frac{V}{T}|w_j(T) - w_j(0)| + \frac{VM}{T}\|z(T) - z(0)\| \quad j \in \{1, \ldots, J\}, \quad (22)$$

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

Proof: We have from the Lipschitz property (3):

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

The first term on the right side above satisfies (20). The second term can be bounded above by $M\frac{V}{T}\|z(T) - z(0)\|$ using (19). This proves the first part of the theorem.

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

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

The first term on the right side above satisfies (18). The second term can be upper bounded by $M\frac{V}{T}\|z(T) - z(0)\|$ using (19). This proves the second part of the theorem. ■

Theorem 2 with Lemma 4 can be interpreted as follows. When $V \geq 1$, the deviation from optimality (21) is bounded above by $O(V/T + 1/V)$, and the constraint violation 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 2 is $O(1/\epsilon^2)$. In fact, this convergence time is in the sense of an one-shot optimization problem. The next section categorizes states of Algorithm 2 as transient phase and steady state phase and analyzes convergence times accordingly.

IV. CONVERGENCE OF TRANSIENT AND STEADY STATE PHASES

Algorithm 2 has two phases: transient phase and steady state phase. Conceptually, we define a steady state by a set of dual variables near optimal Lagrange multipliers of dual problem (11). The transient phase is defined as the period before a generated dual variables arrives at that set. With this idea, we analyze convergence time in two cases of dual function (10) satisfying *locally-polyhedron* and *locally-smooth* properties under the following mild assumption.

Assumption 1: The dual formulation (11) 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., [12], [15]. In addition, Section V shows that the convergences times derived in this section still hold without this uniqueness assumption.

We first provide a general result that will be used later.

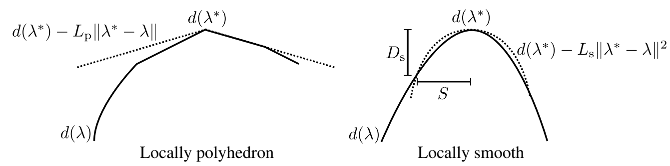

Fig. 1. Illustration of locally-polyhedron and locally-smooth functions

Lemma 5: Let $\{\lambda(t)\}_{t=0}^{\infty}$ be a sequence generated by Algorithm 2. The following relation holds:

$$\|\lambda(t+1) - \lambda^*\|^2 \leq \|\lambda(t) - \lambda^*\|^2 + \frac{2}{V}[d(\lambda(t)) - d(\lambda^*)] + \frac{2C_3}{V^2}, \quad t \in \{0, 1, 2, \ldots\}. \quad (23)$$

Proof: Recall that $\lambda(t) = (w(t), z(t)), h(t) = (g(y(t)), x(t) - y(t))$. From the non-expansive property, we have that

$$\begin{aligned}
&\|\lambda(t+1) - \lambda^*\|^2 \\
&= \left\|\left(\left[w(t) + \tfrac{1}{V}g(y(t))\right]_+, z(t) + \tfrac{1}{V}[x(t) - y(t)]\right) - \lambda^*\right\|^2 \\
&\leq \left\|\left(w(t) + \tfrac{1}{V}g(y(t)), z(t) + \tfrac{1}{V}[x(t) - y(t)]\right) - \lambda^*\right\|^2 \\
&= \|\lambda(t) - \lambda^*\|^2 + \tfrac{1}{V^2}\|h(t)\|^2 + \tfrac{2}{V}[\lambda(t) - \lambda^*]^\top h(t) \\
&\leq \|\lambda(t) - \lambda^*\|^2 + \tfrac{2C_3}{V^2} + \tfrac{2}{V}[d(\lambda(t)) - d(\lambda^*)], \quad (24)
\end{aligned}$$

where the last inequality uses the definition of C_3 and the concavity of the dual function (10), 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-Polyhedron Dual Function

Throughout Section IV-A, the dual function (10) is assumed to have locally-polyhedron property as stated in Assumption 2. This property illustrated in Figure 1.

Assumption 2: Let λ^* be the unique Lagrange multiplier, there exists $L_{\mathrm{p}} > 0$ such that the dual function (10) satisfies

$$d(\lambda^*) \geq d(\lambda) + L_{\mathrm{p}}\|\lambda - \lambda^*\| \quad \text{for all } \lambda \in \Pi. \quad (25)$$

Originally, the locally polyhedron property is defined only for λ near the Lagrange multiplier, but the concavity of dual function (10) implies the property holds for any $\lambda \in \Pi$ as in Figure 1.

The behavior of the generated dual variables with dual function satisfying the locally-polyhedron assumption can be described as follows. Define $B_{\mathrm{p}}(V) \triangleq \max\left\{\frac{L_{\mathrm{p}}}{2V}, \frac{2C_3}{VL_{\mathrm{p}}}\right\}$.

Lemma 6: Under Assumptions 1 and 2, whenever $\|\lambda(t) - \lambda^*\| > B_{\mathrm{p}}(V)$, it follows that

$$\|\lambda(t+1) - \lambda^*\| - \|\lambda(t) - \lambda^*\| < -\frac{L_{\mathrm{p}}}{2V}. \quad (26)$$

Proof: From Lemma 5, suppose the following condition holds

$$\frac{2C_3}{V^2} + \frac{2}{V}[d(\lambda(t)) - d(\lambda^*)] < -\frac{L_{\mathrm{p}}}{V}\|\lambda(t) - \lambda^*\| + \frac{L_{\mathrm{p}}^2}{4V^2}, \quad (27)$$

then inequality (23) becomes

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

Moreover, if $\|\lambda(t)-\lambda^*\| \geq \frac{L_{\rm p}}{2V}$, then the desired inequality (26) holds.

It requires to show that condition (27) holds when $\|\lambda(t)-\lambda^*\| > \frac{2C_3}{VL_{\rm p}}$. However, condition (27) holds when

$$d(\lambda(t)) - d(\lambda^*) < -\frac{C_3}{V} - \frac{L_{\rm p}}{2}\|\lambda(t)-\lambda^*\|.$$

By the locally-polyhedron property (25), if $-L_{\rm p}\|\lambda(t)-\lambda^*\| < -\frac{C_3}{V} - \frac{L_{\rm p}}{2}\|\lambda(t)-\lambda^*\|$, then the above inequality holds. This means that condition (27) holds when $\|\lambda(t)-\lambda^*\| > \frac{2C_3}{VL_{\rm p}}$. This proves the lemma. ∎

Lemma 6 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. The steady state of Algorithm 2 is also defined from this set. This convergence set is defined as

$$\mathcal{R}_{\rm p}(V) = \left\{\lambda \in \Pi : \|\lambda-\lambda^*\| \leq B_{\rm p}(V) + \frac{\sqrt{2C_3}}{V}\right\}. \quad (28)$$

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

$$T_{\rm p} = \arg\inf_{t\geq 0}\{\lambda(t) \in \mathcal{R}_{\rm p}(V)\}. \quad (29)$$

Intuitively, $T_{\rm p}$ is the end of the transient phase and is the beginning of the steady state phase. It is easy to see that $T_{\rm p}$ is at most $O(V)$ from (26).

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

Lemma 7: Under Assumptions 1 and 2, the generated dual variables from Algorithm 2 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 its definition. Suppose that $\lambda(t) \in \mathcal{R}_{\rm p}(V)$. Then two cases are considered. i) If $\|\lambda(t)-\lambda^*\| > B_{\rm p}(V)$, it follows from (26) that $\|\lambda(t+1)-\lambda^*\| < \|\lambda(t)-\lambda^*\| - \frac{L_{\rm p}}{2V} < B_{\rm p}(V) + \frac{\sqrt{2C_3}}{V}$. ii) If $\|\lambda(t)-\lambda^*\| \leq B_{\rm p}(V)$, it follows from the triangle inequality that $\|\lambda(t+1)-\lambda^*\| \leq \|\lambda(t+1)-\lambda(t)\| + \|\lambda(t)-\lambda^*\| \leq \frac{\sqrt{2C_3}}{V} + B_{\rm p}(V)$. 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 3: Under Assumptions 1 and 2, for $T > 0$, let $\{x(t), w(t)\}_{t=T_{\rm p}}^{\infty}$ be a subsequence generated by Algorithm 2, where $T_{\rm p}$ is defined in (29). The following bounds hold:

$$f(\overline{x_{T_{\rm p}}}(T)) - f^{(\rm opt)} \leq \frac{C_3}{V} + \frac{2VM}{T}\left[\frac{\sqrt{2C_3}}{V} + B_{\rm p}(V)\right] + \frac{V}{2T}\left\{\left[\frac{\sqrt{2C_3}}{V} + B_{\rm p}(V)\right]^2 + 4\|\lambda^*\|\left[\frac{\sqrt{2C_3}}{V} + B_{\rm p}(V)\right]\right\} \quad (30)$$

$$g_j(\overline{x_{T_{\rm p}}}(T)) \leq \frac{2V(1+M)}{T}\left[\frac{\sqrt{2C_3}}{V} + B_{\rm p}(V)\right], \quad j \in \{1,\dots,J\}. \quad (31)$$

Proof: The first part of the theorem follows from (21) with the average starting from $T_{\rm p}$ that

$$f(\overline{x_{T_{\rm p}}}(T)) - f^{(\rm opt)} \leq \frac{C_3}{V} + \frac{V}{2T}\left[\|\lambda(T_{\rm p})\|^2 - \|\lambda(T_{\rm p}+T)\|^2\right] + \frac{VM}{T}\|z(T_{\rm p}+T) - z(T_{\rm p})\|. \quad (32)$$

For any $\lambda \in \Pi$, it follows that

$$\|\lambda\|^2 = \|\lambda-\lambda^*\|^2 + \|\lambda^*\|^2 + 2[\lambda-\lambda^*]^\top\lambda^*. \quad (33)$$

The second term on the right-hand-side of (32) can be upper bounded by applying the above equality on $\lambda(T_{\rm p})$ and $\lambda(T_{\rm p}+T)$, i.e.,

$$\begin{aligned}&\|\lambda(T_{\rm p})\|^2 - \|\lambda(T_{\rm p}+T)\|^2 \\ &\leq \|\lambda(T_{\rm p})-\lambda^*\|^2 + 2[\lambda(T_{\rm p}) - \lambda(T_{\rm p}+T)]^\top\lambda^* \\ &\leq \|\lambda(T_{\rm p})-\lambda^*\|^2 + 2\|\lambda(T_{\rm p}) - \lambda(T_{\rm p}+T)\|\|\lambda^*\| \end{aligned} \quad (34)$$

From Lemma 7, the first term of (34) is bounded by $\|\lambda(T_{\rm p})-\lambda^*\|^2 \leq [\sqrt{2C_3}/V + B_{\rm p}(V)]^2$. From triangle inequality and Lemma 7, the last term of (34) is bounded by

$$\|\lambda(T_{\rm p}+T) - \lambda(T_{\rm p})\| \leq \|\lambda(T_{\rm p}+T)-\lambda^*\| + \|\lambda^* - \lambda(T_{\rm p})\| \leq 2\left[\sqrt{2C_3}/V + B_{\rm p}(V)\right]. \quad (35)$$

Therefore, inequality (34) is bounded from above by $[\sqrt{2C_3}/V + B_{\rm p}(V)]^2 + 4\|\lambda^*\|[\sqrt{2C_3}/V + B_{\rm p}(V)]$. Substituting this bound into (32) and using the fact that $\|z(T_{\rm p}+T) - z(T_{\rm p})\| \leq \|\lambda(T_{\rm p}+T) - \lambda(T_{\rm p})\| \leq 2[\sqrt{2C_3}/V + B_{\rm p}(V)]$ proves the first part of the theorem.

The last part follows from (22) that

$$g_j(\overline{x_{T_{\rm p}}}(T)) \leq \frac{V}{T}|w_j(T_{\rm p}+T) - w_j(T_{\rm p})| + \frac{VM}{T}\|z(T_{\rm p}+T) - z(T_{\rm p})\|.$$

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

$$g_j(\overline{x_{T_{\rm p}}}(T)) \leq \frac{V(1+M)}{T}\|\lambda(T_{\rm p}+T) - \lambda(T_{\rm p})\| \leq \frac{2V(1+M)}{T}\left[\frac{\sqrt{2C_3}}{V} + B_{\rm p}(V)\right],$$

where the last inequality uses relation (35). This proves the last part of the theorem. ∎

Theorem 3 can be interpreted as follows. The deviation from the optimality value (30) is bounded above by $O(1/V+1/T)$. The constraint violation (31) 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 2 is $O(1/\epsilon)$. Note that both bounds consider the average starting after reaching the steady state at time T_{p}, and this transient time T_{p} is at most $O(1/\epsilon)$.

B. Locally-Smooth Dual Function

Throughout Section IV-B, the dual function (10) is assumed to have locally-smooth property as stated in Assumption 3. The property is illustrated in Figure 1.

Assumption 3: Let λ^* be the unique Largrange multiplier, there exist $S>0$ and $L_{\text{s}}>0$ such that whenever $\lambda\in\Pi$ and $\|\lambda-\lambda^*\|\leq S$, dual function (10) satisfies

$$d(\lambda^*)\geq d(\lambda)+L_{\text{s}}\|\lambda-\lambda^*\|^2. \tag{36}$$

In addition, there exists $D_s>0$ such that whenever $\lambda\in\Pi$ and $d(\lambda^*)-d(\lambda)\leq D_s$, dual variable satisfies $\|\lambda-\lambda^*\|\leq S$.

The following lemma bounds the order of iteration to get to dual variables that satisfies the above assumption. Note that this result is proven in [13] for a convex function.

Lemma 8: Let $\{\lambda(t)\}_{t=0}^{\infty}$ be the sequence generated by Algorithm 2. Under Assumption 1, for any $\delta>0$, the following holds

$$d(\lambda^*)-\max_{0\leq t\leq E_\delta(V)} d(\lambda(t))\leq\frac{C_3}{V}+\frac{\delta}{2}, \tag{37}$$

where $E_\delta(V)\triangleq\left\lfloor\frac{V\|\lambda(0)-\lambda^*\|^2}{\delta}\right\rfloor$.

Proof:

We prove this lemma by contradiction. Suppose inequality (37) does not hold for all $0\leq t\leq E_\delta(V)$, i.e., $d(\lambda^*)-\max_{0\leq t\leq E_\delta(V)} d(\lambda(t))>\frac{C_3}{V}+\frac{\delta}{2}$. From inequality (23), it follows that for $0\leq t\leq E_\delta(V)$

$$\begin{aligned}\|\lambda(t+1)-\lambda^*\|^2&\leq\|\lambda(t)-\lambda^*\|^2+\frac{2C_3}{V^2}-\frac{2}{V}\left(\frac{C_3}{V}+\frac{\delta}{2}\right)\\&\leq\|\lambda(t)-\lambda^*\|^2-\frac{\delta}{V}.\end{aligned}$$

Summing from $t=0,\dots,E_\delta(V)$ yields:

$$\|\lambda(E_\delta(V)+1)-\lambda^*\|^2\leq\|\lambda(0)-\lambda^*\|^2-\frac{[E_\delta(V)+1]\delta}{V},$$

and $E_\delta(V)+1\leq\frac{V\|\lambda(0)-\lambda^*\|^2}{\delta}$. This contradicts the definition of $E_\delta(V)$. ∎

From Lemma 8, it is easy to see that, when $\delta=D_{\text{s}}$ and $V>2C_3/D_{\text{s}}$, we have $d(\lambda^*)-d(\lambda(\tau))\leq D_{\text{s}}$ for some $0\leq\tau\leq E_\delta(V)$. Then from Assumption 3, we have $\|\lambda(\tau)-\lambda^*\|\leq S$. Thus, by the definition of $E_\delta(V)$, it takes at most $O(V)$ to arrive where the locally-smooth assumption holds.

Then we show the behavior of dual variables, generated by Algorithm 2, that satisfy the locally-smooth assumption. We define $B_{\text{s}}(V)\triangleq\max\left\{\frac{1}{V^{1.5}},\frac{\sqrt{V}+\sqrt{V+4L_{\text{s}}C_3V}}{2L_{\text{s}}V}\right\}$. Then the behavior is stated in the following lemma.

Lemma 9: Under Assumptions 1 and 3, for sufficiently large V that $B_{\text{s}}(V)<S$, whenever $B_{\text{s}}(V)<\|\lambda(t)-\lambda^*\|<S$, it follows that

$$\|\lambda(t+1)-\lambda^*\|-\|\lambda(t)-\lambda^*\|<-\frac{1}{V^{1.5}}. \tag{38}$$

Proof: From Lemma 5, suppose the following condition holds

$$\frac{2C_3}{V^2}+\frac{2}{V}[d(\lambda(t))-d(\lambda^*)]<-\frac{2}{V^{1.5}}\|\lambda(t)-\lambda^*\|+\frac{1}{V^3}, \tag{39}$$

then inequality (23) becomes

$$\begin{aligned}\|\lambda(t+1)-\lambda^*\|^2&<\|\lambda(t)-\lambda^*\|^2-\frac{2}{V^{1.5}}\|\lambda(t)-\lambda^*\|+\frac{1}{V^3}\\&=\left[\|\lambda(t)-\lambda^*\|-\frac{1}{V^{1.5}}\right]^2.\end{aligned}$$

Furthermore, if $\|\lambda(t)-\lambda^*\|\geq\frac{1}{V^{1.5}}$, then the desired inequality (38) holds.

It requires to show that condition (39) holds when $S>\|\lambda(t)-\lambda^*\|>\frac{\sqrt{V}+\sqrt{V+4L_{\text{s}}C_3V}}{2L_{\text{s}}V}$. However, condition (39) holds when

$$d(\lambda(t))-d(\lambda^*)<-\frac{C_3}{V}-\frac{1}{\sqrt{V}}\|\lambda(t)-\lambda^*\|.$$

By the locally-smooth property (36), if $-L_{\text{s}}\|\lambda(t)-\lambda^*\|^2<-\frac{C_3}{V}-\frac{1}{\sqrt{V}}\|\lambda(t)-\lambda^*\|$, then the above inequality holds. This means that condition (39) holds when

$$L_{\text{s}}\|\lambda(t)-\lambda^*\|^2-\frac{1}{\sqrt{V}}\|\lambda(t)-\lambda^*\|-\frac{C_3}{V}>0.$$

The above inequality happens when

$$\begin{aligned}\|\lambda(t)-\lambda^*\|&>\frac{\frac{1}{\sqrt{V}}+\sqrt{\frac{1}{V}+4L_{\text{s}}\frac{C_3}{V}}}{2L_{\text{s}}}\\&=\frac{\sqrt{V}+\sqrt{V+4L_{\text{s}}C_3V}}{2L_{\text{s}}V}.\end{aligned}$$

This prove the lemma. ∎

The interpretation of Lemma 9 is that when the distance between $\lambda(t)$ and λ^* is at least $B_{\text{s}}(V)$ and at most S, then the successor $\lambda(t+1)$ will be closer to λ^*.

Lemma 9 also suggests the existence of a convergence set. The steady state of Algorithm 2 is also defined from this set as

$$\mathcal{R}_{\text{s}}(V)=\left\{\lambda\in\Pi:\|\lambda-\lambda^*\|\leq B_{\text{s}}(V)+\frac{\sqrt{2C_3}}{V}\right\}. \tag{40}$$

Let T_{s} denote the first iteration that a generated dual variables arrives at the convergence set:

$$T_{\text{s}}=\arg\inf_{t\geq 0}\{\lambda(t)\in\mathcal{R}_{\text{s}}(V)\}. \tag{41}$$

It is easy to see that, T_{s} is at most $O(V+V^{1.5})$ from Lemma 8 and (38). Thus, the transient time is at most $O(V^{1.5})$. Next we show that, once the sequence of dual variables enters $\mathcal{R}_{\text{s}}(V)$, it never leaves the set.

Lemma 10: Under Assumptions 1 and 3, for sufficiently large V that $B_{\text{s}}(V)+\frac{\sqrt{2C_3}}{V}<S$, the generated dual variables from Algorithm 2 satisfy $\lambda(t)\in\mathcal{R}_{\text{s}}(V)$ for all $t\geq T_{\text{s}}$.

Proof: We prove the lemma by induction. First we note that $\lambda(T_s) \in \mathcal{R}_s(V)$ by its definition. Suppose that $\lambda(t) \in \mathcal{R}_s(V)$. Then two cases are considered. i) If $\|\lambda(t) - \lambda^*\| > B_s(V)$, it follows from (38) that $\|\lambda(t+1) - \lambda^*\| < \|\lambda(t) - \lambda^*\| - \frac{1}{V^{1.5}} < B_s(V) + \frac{\sqrt{2C_3}}{V}$. ii) If $\|\lambda(t) - \lambda^*\| \leq B_s(V)$, it follows from the triangle inequality that $\|\lambda(t+1) - \lambda^*\| \leq \|\lambda(t+1) - \lambda(t)\| + \|\lambda(t) - \lambda^*\| \leq \frac{\sqrt{2C_3}}{V} + B_s(V)$. Hence, $\lambda(t+1) \in \mathcal{R}_s(V)$ in both cases. This proves the lemma by induction. ∎

Now a convergence of a steady state is ready to be stated.

Theorem 4: Under Assumptions 1 and 3, for sufficiently large V that $B_s(V) + \frac{\sqrt{2C_3}}{V} < S$, for $T > 0$, let $\{x(t), w(t)\}_{t=T_s}^{\infty}$ be a subsequence generated by Algorithm 2, where T_s is defined in (41). The following bounds hold:

$$f(\overline{x_{T_s}}(T)) - f^{(\text{opt})} \leq \frac{C_3}{V} + \frac{2VM}{T}\left[\frac{\sqrt{2C_3}}{V} + B_s(V)\right] + \frac{V}{2T}\left\{\left[\frac{\sqrt{2C_3}}{V} + B_s(V)\right]^2 + 2\|\lambda^*\|\left[\frac{\sqrt{2C_3}}{V} + B_s(V)\right]\right\} \quad (42)$$

$$g_j(\overline{x_{T_s}}(T)) \leq \frac{2V(1+M)}{T}\left[\frac{\sqrt{2C_3}}{V} + B_s(V)\right], \quad j \in \{1, \dots, J\}. \quad (43)$$

Proof: The first part of the theorem follows from (21) with the average starting from T_s that

$$f(\overline{x_{T_s}}(T)) - f^{(\text{opt})} \leq \frac{C_3}{V} + \frac{V}{2T}\left[\|\lambda(T_s)\|^2 - \|\lambda(T_s+T)\|^2\right] + \frac{VM}{T}\|z(T_s+T) - z(T_s)\|. \quad (44)$$

The second term on the right-hand-side of (44) can be bounded from above by applying (33) on $\lambda(T_s)$ and $\lambda(T_s+T)$, i.e.,

$$\begin{aligned}
&\|\lambda(T_s)\|^2 - \|\lambda(T_s+T)\|^2 \\
&\leq \|\lambda(T_s) - \lambda^*\|^2 + 2[\lambda(T_s) - \lambda(T_s+T)]^\top \lambda^* \\
&\leq \|\lambda(T_s) - \lambda^*\|^2 + 2\|\lambda(T_s) - \lambda(T_s+T)\|\|\lambda^*\| \\
&\leq \left[\frac{\sqrt{2C_3}}{V} + B_s(V)\right]^2 + 4\|\lambda^*\|\left[\frac{\sqrt{2C_3}}{V} + B_s(V)\right], \quad (45)
\end{aligned}$$

where the last inequality follows from Lemma 10 and the triangle inequality (similar to (35)).

The last term on the right-hand-side of (44) can be bounded from above by

$$\|z(T_s+T) - z(T_s)\| \leq 2\left[\sqrt{2C_3}/V + B_s(V)\right]. \quad (46)$$

Substituting bounds (45) and (46) into (44) proves the first part of the theorem.

The last part follows from (22) that

$$g_j(\overline{x_{T_s}}(T)) \leq \frac{V}{T}|w_j(T_s+T) - w_j(T_s)| + \frac{VM}{T}\|z(T_s+T) - z(T_s)\|.$$

TABLE I
CONVERGENCE TIMES

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

Since $|w_j(T_s+T) - w_j(T_s)|$ and $\|z(T_s+T) - z(T_s)\|$ are bounded above by $\|\lambda(T_s+T) - \lambda(T_s)\|$, the above inequality is upper bounded by

$$\begin{aligned}
g_j(\overline{x_{T_s}}(T)) &\leq \frac{V(1+M)}{T}\|\lambda(T_s+T) - \lambda(T_s)\| \\
&\leq \frac{2V(1+M)}{T}\left[\frac{\sqrt{2C_3}}{V} + B_s(V)\right].
\end{aligned}$$

This proves the last part of the theorem. ∎

Theorem 4 can be interpreted as follows. The deviation from the optimality (42) is bounded above by $O(1/V + \sqrt{V}/T)$. The constraint violation (43) 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 2 is $O(1/\epsilon^{1.5})$. Note that both bounds consider the average starting after reaching the steady state at time T_s, and this transient time T_s is at most $O(1/\epsilon^{1.5})$.

C. Summary of Convergence Results

The results in Theorems 2, 3, and 4 (denoted by General, Polyhedron, and Smooth) are summarized in Table I. Note that the general convergence time does not have the transient phase and is considered to be in the steady state from the beginning.

V. SAMPLE PROBLEMS

This section illustrates the convergence times of the time-average Algorithm 2 under locally-polyhedron and locally-smooth assumptions. A considered formulation is

$$\begin{aligned}
\text{Minimize} \quad & f(\bar{x}) \qquad (47)\\
\text{Subject to} \quad & 2\bar{x}_1 + \bar{x}_2 \geq 1.5, \quad \bar{x}_1 + 2\bar{x}_2 \geq 1.5 \\
& x_1(t), x_2(t) \in \{0,1,2,3\}, \quad t \in \{0,1,2,\dots\}
\end{aligned}$$

where function f will be given for different cases.

A. 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.

B. Results

Under locally-polyhedron assumption, let $f(x) = 1.5x_1 + x_2$ be the objective function of problem (47). In this setting, the optimal value is 1.25 where $\bar{x}_1 = \bar{x}_2 = 0.5$. Figure 2

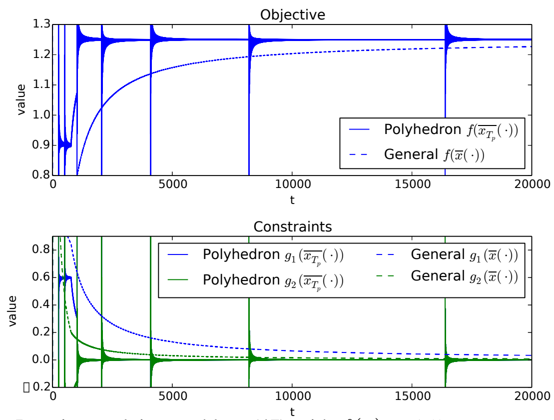

Fig. 2. Iterations solving problem (47) with $f(x) = 1.5x_1 + x_2$

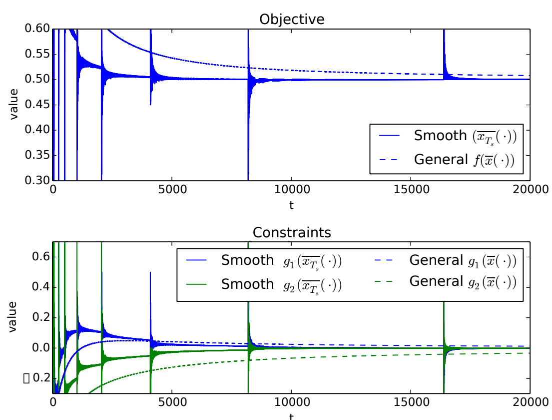

Fig. 3. Iterations solving problem (47) with $f(x) = x_1^2 + x_2^2$

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 polyhedron result ($T_{\text{p}} = 2048$) compared to a general result with convergence time $O(1/\epsilon^2)$.

Under locally-smooth assumption, let $f(x) = x_1^2 + x_2^2$ be the objective function of problem (47). 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 different between convergence times $O(1/\epsilon^2)$ and $O(1/\epsilon^{1.5})$.

Figure 4 illustrates the convergence time of a problem, defined in the figure's caption, without the uniqueness assumption. The Comparison of Figures 4 and 2 shows that there is no difference in the order of convergence time.

VI. CONCLUSION

We consider the time-average optimization problem with a nonconvex (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 $O(1/\epsilon^2)$ of the algorithm that solves the time-average optimization. Under the uniqueness assumption, we prove that faster convergence times $O(1/\epsilon)$ and $O(1/\epsilon^{1.5})$ can be achieved when the average

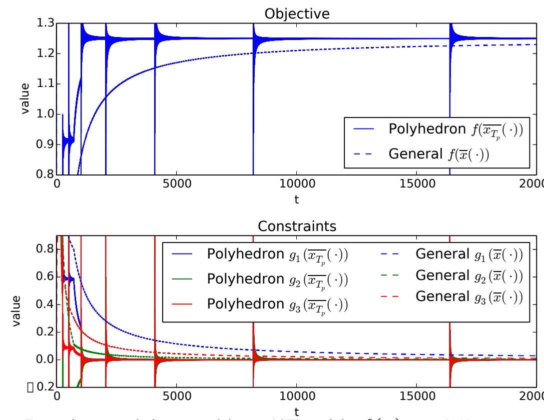

Fig. 4. Iterations solving problem (47) with $f(x) = 1.5x_1 + x_2$ and an additional constraint $\bar{x}_1 + \bar{x}_2 \geq 1$

is performed in the steady state of the algorithm. Then we illustrate by an example that faster convergence time still holds without the uniqueness assumption.

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

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

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

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

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

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

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

[9] A. Beck, A. Nedić, A. Ozdaglar, and M. Teboulle, "Optimal distributed gradient methods for network resource allocation problems," *to appear in IEEE Transactions on Control of Network Systems*, 2013.

[10] E. Wei and A. Ozdaglar, "On the o(1/k) convergence of asynchronous distributed alternating direction method of multipliers," *arXiv:1307.8254*, Jul. 2013.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.