Max-Weight Revisited: Sequences of Non-Convex Optimisations Solving Convex Optimisations

Víctor Valls, Douglas J. Leith
Hamilton Institute, NUI Maynooth

***Abstract*—We investigate the connections between max-weight approaches and dual subgradient methods for convex optimisation. We find that strong connections exist and we establish a clean, unifying theoretical framework that includes both max-weight and dual subgradient approaches as special cases.**

***Index Terms*—convex optimisation, max-weight scheduling, backpressure, subgradient methods**

I. Introduction

In queueing networks, max-weight (also referred to as backpressure) approaches have been the subject of much interest for solving utility optimisation problems in a distributed manner.

In brief, consider a queueing network where the queue occupancy of the i'th queue at time k is denoted by $Q_k^{(i)} \in \mathbb{N}$, $i = 1, 2, \dots, n$, and we gather these together into vector $\boldsymbol{Q}_k \in \mathbb{N}^n$. Time is slotted and at each time step $k = 1, 2, \dots$ we select action $\boldsymbol{x}_k \in D \subset \mathbb{N}^n$, *e.g.*, selecting i'th element $x_k^{(i)} = 1$ corresponds to transmitting one packet from queue i and $x_k^{(i)} = 0$ to doing nothing. The connectivity between queues is captured via matrix $\boldsymbol{A} \in \{-1, 0, 1\}^{n \times n}$, whose i'th row has a -1 at the i'th entry, 1 at entries corresponding to queues from which packets are sent to queue i, and 0 entries elsewhere. The queue occupancy then updates according to $\boldsymbol{Q}_{k+1} = [\boldsymbol{Q}_k + \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{b}_k]^+$, $i = 1, 2, \dots, n$, where the i'th element of vector $\boldsymbol{b}_k \in \mathbb{N}^n$ denotes the number of external packet arrivals to queue i at time k. The objective is to stabilise all of the queues while maximising utility $U(\boldsymbol{z}_k)$ where $U : \mathbb{R}^n \to \mathbb{R}$ is concave and continuously differentiable and $\boldsymbol{z}_k$ is a running average of $\boldsymbol{x}_j$, $j = 1, \dots, k$. The *greedy primal-dual* variant of max-weight scheduling [1], for example, selects action $\boldsymbol{x}_k \in \arg\max_{\boldsymbol{x} \in D} \partial U(\boldsymbol{z}_k)^T \boldsymbol{x} - \beta \boldsymbol{Q}_k^T \boldsymbol{A}\boldsymbol{x}$ with $\boldsymbol{z}_{k+1} = (1-\beta)\boldsymbol{z}_k + \beta \boldsymbol{x}_k$, $0 < \beta < 1$ a design parameter.

Appealing features of this max-weight scheduling approach include the lack of a requirement for *a priori* knowledge of packet arrival process $\{\boldsymbol{b}_k\}$, and the fact that the discrete action set matches the actual decision variables (namely, do we transmit a packet or not). Importantly, although cost function $-U(\cdot)$ is required to be convex, at each time step the max-weight optimisation is non-convex owing to the non-convexity of action set D. Further, convergence is typically proved using Foster-Lyapunov or by sophisticated fluid-limit arguments, which allow sequence $\{\boldsymbol{b}_k\}$ to be accommodated but are distinct from the usual approaches employed in convex optimisation. Hence, the body of work on max-weight approaches remains separate from the mainstream literature on convex optimisation. On the other hand, queueing and Lagrange multiplier subgradient updates are clearly similar, at least superficially, although the exact nature of the relationship between queues and multipliers remains unclear.

Taking these observations as our starting point, in this paper we investigate the connections between max-weight approaches and dual subgradient methods for convex optimisation. We find that strong connections do indeed exist and we establish a clean, unifying theoretical framework that includes both max-weight and dual subgradient approaches as special cases. In summary, the main contributions of the paper include the following.

1) *Generalising max-weight*. Our analysis places max-weight firmly within the field of convex optimisation, extending it from the specific constraints induced by queueing networks to general convex nonlinear contraints with bounded curvature. We show that any non-convex update with suitable descent properties can be employed, and the wealth of convex descent methods can be leveraged to derive non-convex approaches. Descent methods studied here include non-convex variants of the classical Frank-Wolfe update and of the primal Lagrangian update.

2) *Generalising dual subgradient methods*. We show that convexity can be relaxed in classical dual subgradient methods, allowing use of a finite action set. In the special case of optimisation problems with linear constraints, we rigorously establish a close connection (essentially an equivalence) between Lagrange multiplier subgradient updates and discrete queues, so putting existing intuition on a sound footing.

3) *Unifying theoretical framework*. In generalising max-weight and dual subgradient methods our analysis clarifies the fundamental properties required. In particular, bounded curvature of the objective and constraint functions plays a prominent role in our analysis, as does boundedness of the action set. Of interest in its own right, we note that our analysis requires only elementary methods and so an additional contribution is the accessible nature of the methods of proof employed. In particular, it turns out that deterministic analysis of sample paths is sufficient to handle stochasticity. The methods of proof themselves are new in the context of max-weight approaches, and are neither Foster-Lyapunov nor fluid-limit based.

A. Related Work

Max-weight scheduling was introduced by Tassiulas and Ephremides in their seminal paper [2]. They consider a

This work was supported by Science Foundation Ireland under Grant No. 11/PI/1177.

network of queues with slotted time, an integer number of packet arrivals in each slot and a finite set of admissible scheduling patterns, referred to as *actions*, in each slot. Using a Forster-Lyapunov approach they present a scheduling policy that stabilises the queues provided the external traffic arrivals are strictly feasible. Namely, the scheduling policy consists of selecting the action at each slot that maximises the queue-length-weighted sum of rates, $\boldsymbol{x}_k \in \arg\max_{\boldsymbol{x}\in D} -\boldsymbol{Q}_k^T\boldsymbol{A}\boldsymbol{x}$.

Independently, [1], [3], [4] proposed extensions to the max-weight approach to accommodate concave utility functions. In [1] the *greedy primal-dual* algorithm is introduced, as already described above, for network linear constraints and utility function $U(\cdot)$ which is continuously differentiable and concave. The previous work is extended in [5] to consider general nonlinear constraints. In [4] the utility fair allocation of throughput in a cellular downlink is considered. The utility function is of the form $U(\boldsymbol{z}) = \sum_{i=1}^n U_i(z^{(i)})$, $U_i(z) = \beta_i \frac{z^{(1-\frac{1}{m})}}{1-\frac{1}{m}}$, with m, β_i design parameters. Queue departures are scheduled according to $\boldsymbol{x}_k \in \arg\max_{\boldsymbol{x}\in \mathrm{conv}(D)} -\boldsymbol{Q}_k^T\boldsymbol{A}\boldsymbol{x}$ and queue arrivals are scheduled by a congestion controller such that $E[b_k^{(i)}|\boldsymbol{Q}_k] = \min\{\partial U_i(Q_k^{(i)}), M\}$ and $E[(b_k^{(i)})^2|\boldsymbol{Q}_k] \leq A$ where A, M are positive constants. The work in [3] considers power allocation in a multibeam downlink satellite communication link with the aim of maximising throughput while ensuring queue stability. This is extended in a sequence of papers [6], [7], [8] and a book [9] to develop the *drift plus penalty* approach. In this approach the basic strategy for scheduling queue departures is according to $\boldsymbol{x}_k \in \arg\max_{\boldsymbol{x}\in D} -\boldsymbol{Q}_k^T\boldsymbol{A}\boldsymbol{x}$ and utility functions are incorporated in a variety of ways. For example, for concave non-decreasing continuous utility functions U of the form $U(\boldsymbol{z}) = \sum_{i=1}^n U_i(z^{(i)})$ one formulation is for a congestion controller to schedule arrivals into an ingress queue such that $b_k^{(i)} \in \arg\max_{0\leq b\leq R} VU_i(b) - bQ_k^{(i)}$ where V, R are sufficiently large design parameters and $b \in \mathbb{R}$ [10]. Another example is for cost functions of the form $E[P_k(\boldsymbol{x}_k)]$ where $P_k(\cdot)$ is bounded, i.i.d. and known at each time step, in which case actions at each time step are selected to minimise $\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} VP_k(\boldsymbol{x}_k) + \boldsymbol{Q}_k^T\boldsymbol{A}\boldsymbol{x}$ where V is a design parameter [9].

With regard to the existence of a connection between the discrete-valued queue occupancy in a queueing network and continuous-valued Lagrange multipliers, this has been noted by several authors, see for example [11], [12], and so might be considered something of a "folk theorem" but we are aware of few rigorous results. A notable exception is [13], which establishes that a discrete queue update tends on average to drift towards the optimal multiplier value. Also, the *greedy primal-dual* algorithm presented in [1] shows that asymptotically as design parameter $\beta \to 0$ and $t \to \infty$ the scaled queue occupancy converges to the set of dual optima.

Selection of a sequence of actions in a discrete-like manner is also considered in the convex optimisation literature. The *nonlinear Gauss-Seidel* algorithm, also known as *block coordinate descent* [14], [15] minimises a convex function over a convex set by updating one co-ordinate at a time. The convex function is required to be continuously differentiable and strictly convex and, unlike in the max-weight algorithms discussed above, the action set is convex. The classical Frank-Wolfe algorithm [16] also minimises a convex continuously differentiable function over a polytope by selecting from a discrete set of descent directions, although a continuous-valued line search is used to determine the final update. We also note the work on online convex optimisation [17], [18], where the task is to choose a sequence of actions so to minimise an unknown sequence of convex functions with low regret.

B. Notation

Vectors and matrices are indicated in bold type. Since we often use subscripts to indicate elements in a sequence, to avoid confusion we usually use a superscript $x^{(i)}$ to denote the i'th element of a vector $\boldsymbol{x}$. The i'th element of operator $[\boldsymbol{x}]^{[0,\bar{\lambda})}$ equals $x^{(i)}$ (the i'th element of $\boldsymbol{x}$) when $x^{(i)} \in [0,\bar{\lambda})$ and otherwise equals 0 when $x^{(i)} < 0$ and $\bar{\lambda}$ when $x^{(i)} \geq \bar{\lambda}$. Note that we allow $\bar{\lambda} = +\infty$, and following standard notation in this case usually write $[x]^+$ instead of $[x]^{[0,\infty)}$. The subgradient of a convex function f at point $\boldsymbol{x}$ is denoted $\partial f(\boldsymbol{x})$. For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ we use element-wise comparisons $\boldsymbol{x} \succeq \boldsymbol{y}$ and $\boldsymbol{y} \succ \boldsymbol{x}$ to denote when $y^{(i)} \geq x^{(i)}$, $y^{(i)} > x^{(i)}$ respectively for all $i = 1, \ldots, m$.

II. Preliminaries

We recall the following convexity properties.

Lemma 1 (Lipschitz Continuity). *Let $h : M \to \mathbb{R}$ be a convex function and let C be a closed and bounded set contained in the relative interior of the domain $M \subset \mathbb{R}^n$. Then $h(\cdot)$ is Lipschitz continuous on C* i.e., *there exists constant ν_h such that $|h(\boldsymbol{x}) - h(\boldsymbol{y})| \leq \nu_h\|\boldsymbol{x}-\boldsymbol{y}\|_2$ $\forall \boldsymbol{x}, \boldsymbol{y} \in C$.*

Proof: See, for example, [19]. ∎

Lemma 2 (Bounded Distance). *Let $D := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{|D|}\}$ be a finite set of points from $\mathbb{R}^n$. Then there exists constant $\bar{x}_D$ such that $\|\boldsymbol{z}-\boldsymbol{y}\|_2 \leq \bar{x}_D$ for any two points $\boldsymbol{z}, \boldsymbol{y} \in C := \mathrm{conv}(D)$, where $\mathrm{conv}(D)$ denotes the convex hull of D.*

Proof: Since $\boldsymbol{z}, \boldsymbol{y} \in C$ these can be written as the convex combination of points in D, i.e., $\boldsymbol{z} = \sum_{i=1}^{|D|} a^{(i)}\boldsymbol{x}_i$, $\boldsymbol{y} = \sum_{i=1}^{|D|} b^{(i)}\boldsymbol{x}_i$ with $\|\boldsymbol{a}\|_1 = 1 = \|\boldsymbol{b}\|_1$. Hence $\|\boldsymbol{z}-\boldsymbol{y}\|_2 = \|\sum_{i=1}^{|D|}(a^{(i)}-b^{(i)})\boldsymbol{x}_i\|_2 \leq \sum_{i=1}^{|D|}\|a^{(i)}-b^{(i)}\|_2\|\boldsymbol{x}_i\|_2 \leq \bar{x}_D := 2\max_{\boldsymbol{x}\in D}\|\boldsymbol{x}\|_2$. ∎

We also introduce the following definition:

Definition 1 (Bounded Curvature). *Let $h : M \to \mathbb{R}$ be a convex function defined on domain $M \subset \mathbb{R}^n$. We say the $h(\cdot)$ has bounded curvature on set $C \subset M$ if for any points $\boldsymbol{z}, \boldsymbol{z}+\boldsymbol{\delta} \in C$*

$$h(\boldsymbol{z}+\boldsymbol{\delta}) - h(\boldsymbol{z}) \leq \partial h(\boldsymbol{z})^T\boldsymbol{\delta} + \mu_h\|\boldsymbol{\delta}\|_2^2 \qquad (1)$$

where $\mu_h \geq 0$ is a constant that does not depend on $\boldsymbol{z}$ or $\boldsymbol{\delta}$.

Bounded curvature will prove important in our analysis. The following lemma shows that a necessary and sufficient condition for bounded curvature is that the subgradients of $h(\cdot)$ are Lipschitz continuous on set C.

Lemma 3 (Bounded Curvature)**.** *Let $h : M \to \mathbb{R}$, $M \subset \mathbb{R}^n$ be a convex function. Then $h(\cdot)$ has bounded curvature on C if and only if for all $\boldsymbol{z}, \boldsymbol{z}+\boldsymbol{\delta} \in C$ there exists a member $\partial h(\boldsymbol{z})$ (respectively, $\partial h(\boldsymbol{z}+\boldsymbol{\delta})$) of the set of subdifferentials at point $\boldsymbol{z}$ (respectively, $\boldsymbol{z}+\boldsymbol{\delta}$) such that $\left(\partial h(\boldsymbol{z}+\boldsymbol{\delta}) - \partial h(\boldsymbol{z})\right)^T \boldsymbol{\delta} \leq \mu_h \|\boldsymbol{\delta}\|_2^2$ where μ_h does not depend on $\boldsymbol{z}$ or $\boldsymbol{\delta}$.*

Proof: $\Rightarrow$ Suppose $h(\cdot)$ has bounded curvature on C. From (1) it follows that $h(\boldsymbol{z}+\boldsymbol{\delta}) - h(\boldsymbol{z}) \leq \partial h(\boldsymbol{z})^T\boldsymbol{\delta} + \mu_h\|\boldsymbol{\delta}\|_2^2$ and $h(\boldsymbol{z}) - h(\boldsymbol{z}+\boldsymbol{\delta}) \leq -\partial h(\boldsymbol{z}+\boldsymbol{\delta})^T\boldsymbol{\delta} + \mu_h\|\boldsymbol{\delta}\|_2^2$. Adding left-hand and right-hand sides of these inequalities yields $0 \leq \left(\partial h(\boldsymbol{z}) - \partial h(\boldsymbol{z}+\boldsymbol{\delta})\right)^T \boldsymbol{\delta} + 2\mu_h\|\boldsymbol{\delta}\|_2^2$ i.e., $\left(\partial h(\boldsymbol{z}+\boldsymbol{\delta}) - \partial h(\boldsymbol{z})\right)^T \boldsymbol{\delta} \leq \mu_h\|\boldsymbol{\delta}\|_2^2$.

$\Leftarrow$ Suppose $\left(\partial h(\boldsymbol{z}+\boldsymbol{\delta}) - \partial h(\boldsymbol{z})\right)^T \boldsymbol{\delta} \leq \mu_h\|\boldsymbol{\delta}\|_2$ for all $\boldsymbol{z}, \boldsymbol{z}+\boldsymbol{\delta} \in M$. It follows that $\partial h(\boldsymbol{z}+\boldsymbol{\delta})^T\boldsymbol{\delta} \leq \partial h(\boldsymbol{z})^T\boldsymbol{\delta} + \mu_h\|\boldsymbol{\delta}\|_2^2$. By the definition of the subgradient we have that $h(\boldsymbol{z}+\boldsymbol{\delta}) - h(\boldsymbol{z}) \leq \partial h(\boldsymbol{z}+\boldsymbol{\delta})^T\boldsymbol{\delta}$, and so we obtain that $h(\boldsymbol{z}+\boldsymbol{\delta}) - h(\boldsymbol{z}) \leq \partial h(\boldsymbol{z})^T\boldsymbol{\delta} + \mu_h\|\boldsymbol{\delta}\|_2^2$. ■

III. Non-Convex Descent

We begin by considering minimisation of convex function $F : \mathbb{R}^n \to \mathbb{R}$ on convex set $C := \mathrm{conv}(D)$, the convex hull of set $D := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{|D|}\}$ consisting of a finite collection of points from $\mathbb{R}^n$ (so C is a polytope). Our interest is in selecting a sequence of points $\{\boldsymbol{x}_k\}$, $k = 1, 2, \ldots$ from set D such that the running average $\boldsymbol{z}_{k+1} = (1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}_k$ minimises $F(\cdot)$ for k sufficiently large and β sufficiently small. Note that set D is non-convex since it consists of a finite number of points, and by analogy with max-weight terminology we will refer to it as the *action set*.

Since C is the convex hull of action set D, any point $\boldsymbol{z}^* \in C$ minimising $F(\cdot)$ can be written as convex combinations of points in D i.e., $\boldsymbol{z}^* = \sum_{i=1}^{|D|} a^{*(i)}\boldsymbol{x}_i$, $a^{*(i)} \in [0, 1]$, $\|\boldsymbol{a}^*\|_1 = 1$. Hence, we can always construct sequence $\{\boldsymbol{x}_k\}$ by selecting points from set D in proportion to the $a^{*(i)}$, $i = 1, \ldots, |D|$. That is, by *a posteriori* time-sharing (*a posteriori* in the sense that we need to find minimum $\boldsymbol{z}^*$ before we can construct sequence $\{\boldsymbol{x}_k\}$). Of more interest, however, it turns out that when function $F(\cdot)$ has bounded curvature then sequences $\{\boldsymbol{x}_k\}$ can be found without requiring knowledge of $\boldsymbol{z}^*$.

A. Non-Convex Direct Descent

The following theorem formalises the above commentary, also generalising it to sequences of convex functions $\{F_k\}$ rather than just a single function as this will prove useful later.

Theorem 1 (Greedy Non-Convex Convergence)**.** *Let $\{F_k\}$ be a sequence of convex functions with uniformly bounded curvature μ_F on set $C := \mathrm{conv}(D)$, action set D a finite set of points from $\mathbb{R}^n$. Let $\{\boldsymbol{z}_k\}$ be a sequence of vectors satisfying $\boldsymbol{z}_{k+1} = (1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}_k$ with $\boldsymbol{z}_1 \in C$ and*

$$\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} F_k((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}),\ k = 1, 2, \ldots \tag{2}$$

Suppose parameter β is sufficiently small that

$$0 < \beta \leq (1-\gamma)\min\{\epsilon/(\mu_F\bar{x}_D^2), 1\} \tag{3}$$

with $0 < \gamma < 1$, $\epsilon > 0$ and that functions F_k change sufficiently slowly that

$$|F_{k+1}(\boldsymbol{z}) - F_k(\boldsymbol{z})| \leq \gamma_1\gamma\beta\epsilon,\ \forall \boldsymbol{z} \in C \tag{4}$$

with $0 < \gamma_1 < 1/2$. Then for k sufficiently large we have that

$$0 \leq F_k(\boldsymbol{z}_{k+1}) - F_k(\boldsymbol{y}_k^*) \leq 2\epsilon \tag{5}$$

where $\boldsymbol{y}_k^* \in \arg\min_{\boldsymbol{z}\in C} F_k(\boldsymbol{z})$.

Proof: See Appendix. ■

Observe that in Theorem 1 we select $\boldsymbol{x}_k$ by solving non-convex optimisation (2) at each time step. This optimisation is one step ahead, or greedy, in nature and does not look ahead to future values of the sequence or require knowledge of optimum $\boldsymbol{z}^*$. Of course, such an approach is mainly of interest when non-convex optimisation (2) can be efficiently solved, *e.g.*, when action set D is small or the optimisation separable.

Observe also that Theorem 1 relies upon the bounded curvature of the $F_k(\cdot)$. A smoothness assumption of this sort seems essential, since when it does not hold it is easy to construct examples where Theorem 1 does not hold. Such an example is illustrated schematically in Figure 1a. The shaded region in Figure 1a indicates the level set $\{F(\boldsymbol{y}) \leq F(\boldsymbol{z}) : \boldsymbol{y} \in C\}$. The level set is convex, but the boundary is non-smooth and contains "kinks". We can select points from the set $\{(1-\beta)\boldsymbol{z} + \beta\boldsymbol{x} : \boldsymbol{x} \in D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}\}$. This set of points is indicated in Figure 1a and it can be seen that every point lies outside the level set. Hence, we must have $F((1-\beta)\boldsymbol{z} + \beta\boldsymbol{x}) > F(\boldsymbol{z})$, and upon iterating we will end up with a diverging sequence. Note that in this example changing the step size β does not resolve the issue. Bounded curvature ensures that the boundary of the level sets is smooth, and this ensures that for sufficiently small β there exists a convex combination of $\boldsymbol{z}$ with a point $\boldsymbol{x} \in D$ such that $F((1-\beta)\boldsymbol{z} + \beta\boldsymbol{x}) < F(\boldsymbol{z})$ and so the solution to optimisation (2) improves our objective, see Figure 1b.

Theorem 1 is stated in a fairly general manner since this will be needed for our later analysis. An immediate corollary to Theorem 1 is the following convergence result for unconstrained optimisation.

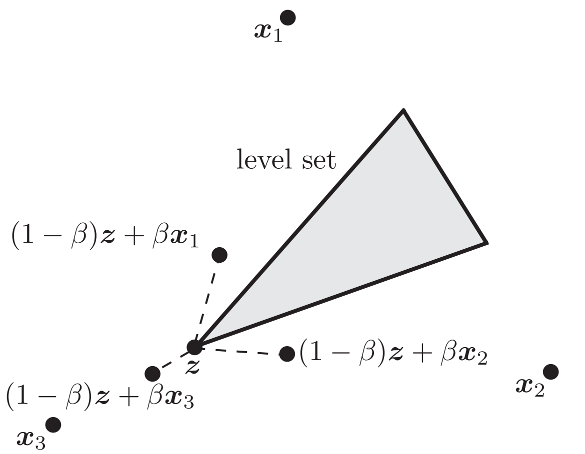
Corollary 1 (Unconstrained Optimisation)**.** *Consider the following sequence of non-convex optimisations $\{P_k^u\}$:*

$$\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} f\left((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}\right)$$
$$\boldsymbol{z}_{k+1} = (1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}_k$$

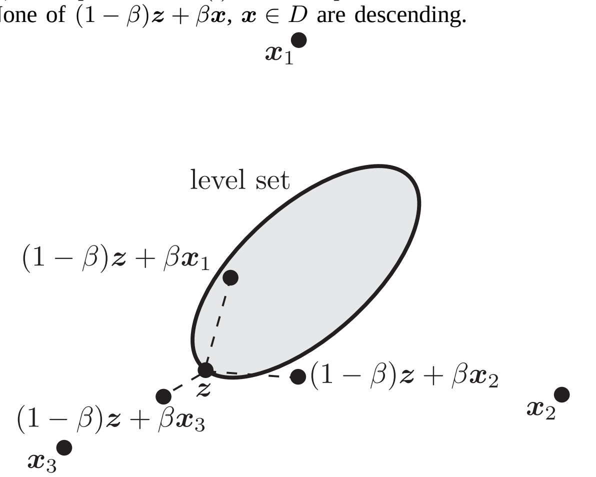
with $\boldsymbol{z}_1 \in C := \mathrm{conv}D$, action set $D \subset \mathbb{R}^n$ finite. Suppose $f(\cdot)$ has bounded curvature with curvature constant μ_f. Then for any $\epsilon > 0$ and $0 < \beta \leq (1-\gamma)\min\{\epsilon/(\mu_f\bar{x}_D^2), 1\}$ with $0 < \gamma < 1$ we have $0 \leq f(\boldsymbol{z}_k) - f^* \leq 2\epsilon$ for all k sufficiently large, where $f^* = \min_{\boldsymbol{z}\in C} f(\boldsymbol{z})$.

Figure 2 illustrates Corollary 1 schematically in $\mathbb{R}^2$. The sequence of non-convex optimisations descends in two iterations $f(\boldsymbol{z}_1) > f(\boldsymbol{z}_2) > f(\boldsymbol{z}_3)$ (using points $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$ respectively) and $f(\boldsymbol{z}_k) - f^* \leq 2\epsilon$ for $k > 3$ (not shown in Figure 2).

When the curvature constant of function f need not be known, an upper bound is sufficient to select β. Next we



(a) Example where $F(\cdot)$ does not possess bounded curvature. None of $(1-\beta)\boldsymbol{z}+\beta\boldsymbol{x}$, $\boldsymbol{x}\in D$ are descending.



(b) Example where $F(\cdot)$ has bounded curvature. For β sufficiently small, for at least one $(1-\beta)\boldsymbol{z}+\beta\boldsymbol{x}$, $\boldsymbol{x}\in D$ descent is possible.

Fig. 1: Illustrating how bounded curvature allows monotonic descent. Set D consists of the marked points $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$. Level set $\{F(\boldsymbol{y})\le F(\boldsymbol{z}):\boldsymbol{y}\in C\}$ is indicated by the shaded areas. The possible choices of $(1-\beta)\boldsymbol{z}+\beta\boldsymbol{x}$, $\boldsymbol{x}\in D$ are indicated.

present two examples that are affected differently by constant μ_f.

Example 1 (Linear Objective)**.** Suppose $f(\boldsymbol{z}) := \boldsymbol{a}^T\boldsymbol{z}$ where $\boldsymbol{a}\in\mathbb{R}^n$. The objective function is linear and so has curvature constant $\mu_f=0$. For any $\beta\in(0,1)$ we have that $f(\boldsymbol{z}_{k+1}) < f(\boldsymbol{z}_k)$ for all k and $f(\boldsymbol{z}_k)\to f^*$.

Example 2 (Quadratic Objective)**.** Suppose $f(\boldsymbol{z}) := \frac{1}{2}\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}$ where $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ is positive definite. Then, if we choose $\mu_f\ge\lambda_{\max}(\boldsymbol{A})$ we have that $f(\boldsymbol{z}_k)-f^*\le 2\epsilon$ for k sufficiently large.

B. Non-Convex Frank-Wolfe-like Descent

It is important to note that other convergent non-convex updates are also possible. For example:

Theorem 2 (Greedy Non-Convex FW Convergence)**.** *Consider the setup in Theorem 1, but with modified update*

$$\boldsymbol{x}_k\in\arg\min_{\boldsymbol{x}\in D}\partial F_k(\boldsymbol{z}_k)^T\boldsymbol{x},\ \ k=1,2,\ldots \tag{6}$$

Then for k sufficiently large we have that

$$0\le F_k(\boldsymbol{z}_{k+1})-F_k(\boldsymbol{y}_k^*)\le 2\epsilon \tag{7}$$

where $\boldsymbol{y}_k^*\in\arg\min_{\boldsymbol{z}\in C}F_k(\boldsymbol{z})$.

Proof: See Appendix. ■

The intuition behind the update in Theorem 2 is that at each step we locally approximate $F_k(\boldsymbol{z}_{k+1})$ by linear function $F_k(\boldsymbol{z}_k)+\partial F_k(\boldsymbol{z}_k)^T(\boldsymbol{z}_{k+1}-\boldsymbol{z}_k)$ and then minimise this linear function. Since $F_k(\cdot)$ is convex, this linear function is in fact the supporting hyperplane to $F_k(\cdot)$ at point $\boldsymbol{z}_k$, and so can be expected to allow us to find a descent direction. Similar intuition also underlies classical Frank-Wolfe algorithms for convex optimisation [16] on a polytope, and Theorem 2 extends this class of algorithms to make use of non-convex update (6) and a fixed step size (rather than the classical approach of selecting the step size by line search).

Note that when the function is linear $F_k(\boldsymbol{z})=\boldsymbol{c}_k^T\boldsymbol{z}$, $\boldsymbol{c}_k\in\mathbb{R}^n$, then $\arg\min_{\boldsymbol{x}\in D}F_k((1-\beta)\boldsymbol{z}+\beta\boldsymbol{x})=\arg\min_{\boldsymbol{x}\in D}\boldsymbol{c}_k^T\boldsymbol{x}$ and $\arg\min_{\boldsymbol{x}\in D}\partial F_k(\boldsymbol{z}_k)^T\boldsymbol{x} = \arg\min_{\boldsymbol{x}\in D}\boldsymbol{c}_k^T\boldsymbol{x}$. That is, updates (2) and (6) are identical.

Note also that

$$\arg\min_{\boldsymbol{x}\in D}\partial F_k(\boldsymbol{z}_k)^T\boldsymbol{x}\subseteq\arg\min_{\boldsymbol{z}\in C}\partial F_k(\boldsymbol{z}_k)^T\boldsymbol{z}. \tag{8}$$

This is because the RHS of (8) is a linear program (the objective is linear and set C is a polytope, so defined by linear constraints) and so the optimum set is either (i) an extreme point of C and so a member of set D, or (ii) a face of polytope C with the extreme points of the face belonging to set D. Hence, while update (6) is non-convex it can nevertheless be solved in polynomial time.

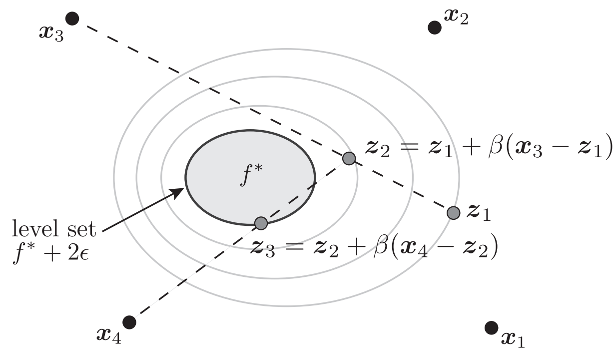


Fig. 2: Illustrating unconstrained convergence in $\mathbb{R}^2$. The sequence of non-convex optimisations converges with $k=2$. The function average decreases monotonically when $f(\boldsymbol{z}_k) > f^*+2\epsilon$, and when $f(\boldsymbol{z}_k)\le f^*+2\epsilon$ we have that $f(\boldsymbol{z}_k)$ remains in the level set for the next iterations.

IV. Queues and Multipliers

Before proceeding to consider solving constrained optimisation problems, in this section we first establish a useful relationship between discrete and continuous queues, and also a key lower bound on the average Lagrangian when using a dual subgradient update.

A. Discrete and Continuous Queues

As before, let sequence $\boldsymbol{x}_k$ consist of points from finite action set $D \subset \mathbb{R}^n$ and $\boldsymbol{z}_{k+1} = (1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}_k$, $0 < \beta < 1$. Let matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and suppose $\{\boldsymbol{b}_k\}$ is a sequence of points from set $E \subseteq \mathbb{R}^m$ such that $\boldsymbol{b} := \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^k \boldsymbol{b}_i$ exists. Set E need not be convex and may consist of a finite number of points *e.g.*, we might have $E = \{0,1\}^m$. Hence, we might think of $\boldsymbol{b}_k$ as the packet arrivals/departures at time step k and $\boldsymbol{b}$ as the mean packet arrival/departure rate. Define the following two queueing iterations:

$$\boldsymbol{Q}_{k+1} = [\boldsymbol{Q}_k + \boldsymbol{A}\boldsymbol{z}_{k+1} - \boldsymbol{b}]^+, \tag{9}$$
$$\tilde{\boldsymbol{Q}}_{k+1} = [\tilde{\boldsymbol{Q}}_k + \boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}_k]^+ \tag{10}$$

with $\boldsymbol{Q}_k$, $\tilde{\boldsymbol{Q}}_k \in \mathbb{R}^m_+$, initial condition $\boldsymbol{Q}_1 = \tilde{\boldsymbol{Q}}_1$. Here, vector $\boldsymbol{Q}_k$ can be thought of as the occupancy of a set of queues with averaged arrivals/departures $\boldsymbol{A}\boldsymbol{z}_{k+1}$ and $\boldsymbol{b}$[1]. While vector $\tilde{\boldsymbol{Q}}_k$ can be thought of as the occupancy of a set of queues with discrete arrivals/departures $\boldsymbol{A}\boldsymbol{x}_k$ and $\boldsymbol{b}_k$.

When $\beta = 1$ and $\boldsymbol{b}_k = \boldsymbol{b}$ then $\boldsymbol{z}_{k+1} = \boldsymbol{x}_k$, the two queue updates are identical and $\boldsymbol{Q}_k = \tilde{\boldsymbol{Q}}_k$. For smaller values of β and for $\boldsymbol{b}_k \neq \boldsymbol{b}$, so long as the differences between $\boldsymbol{z}_{k+1}$, $\boldsymbol{x}_k$ and between $\boldsymbol{b}_k$, $\boldsymbol{b}$ remain small in an appropriate sense then we might expect that the queue occupancies $\boldsymbol{Q}_k$ and $\tilde{\boldsymbol{Q}}_k$ remain close, and indeed this is the case. The required sense in which differences are required to be small is clarified by the following lemma, which is a direct result of [20, Proposition 3.1.2] and we include the proof for completeness.

Lemma 4. *Consider the sequences* $x_{k+1} = [x_k + \delta_k]^+$ *and* $y_{k+1} = [y_k + \tilde{\delta}_k]^+$, $k = 1,2,\dots$ *where* $x_k, y_k \in \mathbb{R}_+$, $\delta_k, \tilde{\delta}_k \in \mathbb{R}$. *Suppose* $x_1 = y_1$ *and* $|\sum_{i=1}^k \delta_i - \tilde{\delta}_i| \le \epsilon$ *for all* $k \ge 1$. *Then,*

$$|x_k - y_k| \le 2\epsilon \qquad k \ge 1.$$

Proof: We proceed by considering two cases:

Case (i) ($x_1, y_1 = 0$): Let $\eta_k := -\min(x_k + \delta_k, 0)$, i.e., $x_{k+1} = x_k + \delta_k + \eta_k$ and so we can write $x_{k+1} = \sum_{i=1}^k \delta_i + \eta_i$. See from the previous equations that if $x_{k+1} > 0$ then $\eta_k = 0$, and that if $x_{k+1} = 0$ then $-\sum_{i=1}^k \delta_i = \sum_{i=1}^k \eta_i$ and $\eta_k \ge 0$. Hence, we have that $\sum_{i=1}^k \eta_i$ is nondecreasing with k and so it must be that $\sum_{i=1}^k \eta_i = -\min_{1\le j\le k}\sum_{i=1}^j \delta_i$. Then, we can write $x_{k+1} = \sum_{i=1}^k \delta_i - \min_{1\le j\le k}\sum_{i=1}^j \delta_i$ and $y_{k+1} = \sum_{i=1}^k \tilde{\delta}_i - \min_{1\le j\le k}\sum_{i=1}^j \tilde{\delta}_i$. Next, see that

$$\begin{aligned}
&|x_{k+1} - y_{k+1}| \\
&= \left|\sum_{i=1}^k \delta_i - \tilde{\delta}_i + \left(\min_{1\le j\le k}\sum_{i=1}^j \tilde{\delta}_j\right) - \left(\min_{1\le j\le k}\sum_{i=1}^j \delta_j\right)\right| \\
&\le \left|\sum_{i=1}^k \delta_i - \tilde{\delta}_i\right| + \left|\left(\max_{1\le j\le k} - \sum_{i=1}^j \tilde{\delta}_j\right) - \left(\max_{1\le j\le k} - \sum_{i=1}^j \delta_j\right)\right| \\
&\le \left|\sum_{i=1}^k \delta_i - \tilde{\delta}_i\right| + \max_{1\le j\le k}\left|\sum_{i=1}^j \delta_j - \sum_{i=1}^j \tilde{\delta}_j\right|
\end{aligned}$$

and since $|\sum_{i=1}^k \delta_i - \tilde{\delta}_i| \le \epsilon$ for all k the desired bound follows.

Case (ii) ($x_1, y_1 > 0$): Since $x_1 > 0$ we now have that $x_{k+1} = x_1 + \sum_{i=1}^k \delta_i + [-\min_{1\le j\le k}\sum_{i=1}^j \delta_i - x_1]^+ = \sum_{i=1}^k \delta_i + \max\{-\min_{1\le j\le k}\sum_{i=1}^j \delta_i, x_1\}$. Next, let $a := -\min_{1\le j\le k}\sum_{i=1}^j \delta_i$, $b := -\min_{1\le j\le k}\sum_{i=1}^j \tilde{\delta}_i$, $c := x_1 = y_1$ and observe that since $a, b, c \ge 0$ we have that $|\max\{a,c\} - \max\{b,c\}| \le |a-b|$. Hence, we have like in the previous case that $|x_{k+1} - y_{k+1}| \le |\sum_{i=1}^k \delta_i - \tilde{\delta}_i| + |a-b|$ and the stated result follows. ■

It can be seen from Lemma 4 that the requirement is that $|\sum_{i=1}^k \delta_i - \tilde{\delta}_i|$ is uniformly bounded for all k. Hence, to bound the difference between queue occupancies $\boldsymbol{Q}_k$ and $\tilde{\boldsymbol{Q}}_k$ we can expect to require that $|\sum_{i=1}^k b_i^{(j)} - b^{(j)}| \le \sigma_2$ for some $\sigma_2 > 0$ *i.e.*, that $|(\frac{1}{k}\sum_{i=1}^k b_i^{(j)}) - b^{(j)}| \le \sigma_2/k$. Also that will need to show that $|\sum_{i=1}^k (\boldsymbol{a}^{(j)})^T(\boldsymbol{x}_k - \boldsymbol{z}_k)|$ is uniformly bounded, where $(\boldsymbol{a}^{(j)})^T$ denotes the j'th row of matrix $\boldsymbol{A}$. However, rather than stating this result in terms of $\boldsymbol{Q}_k$ and $\tilde{\boldsymbol{Q}}_k$ it will prove more convenient later to state it in terms of the corresponding rescaled updates

$$\boldsymbol{\lambda}_{k+1} = [\boldsymbol{\lambda}_k + \alpha(\boldsymbol{A}\boldsymbol{z}_{k+1} - \boldsymbol{b})]^+, \tag{11}$$
$$\tilde{\boldsymbol{\lambda}}_{k+1} = [\tilde{\boldsymbol{\lambda}}_k + \alpha(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}_k)]^+ \tag{12}$$

where $\boldsymbol{\lambda}_k := \alpha\boldsymbol{Q}_k$ and $\tilde{\boldsymbol{\lambda}}_k := \alpha\tilde{\boldsymbol{Q}}_k$, $\alpha > 0$. We have the following:

Lemma 5 (Discrete and Continuous Queues)**.** *Consider the updates (11) and (12) with* $\boldsymbol{\lambda}_1 = \tilde{\boldsymbol{\lambda}}_1$. *Suppose that* $|(\frac{1}{k}\sum_{i=1}^k b_i^{(j)}) - b^{(j)}| \le \sigma_2/k$ *where* $b_i^{(j)}$ *denotes the* j*'th element of vector* $\boldsymbol{b}_i$. *Then,*

$$|\tilde{\lambda}_k^{(j)} - \lambda_k^{(j)}| \le 2\alpha(\sigma_1/\beta + \sigma_2),\ \ k = 1,2,\dots \tag{13}$$

where $\tilde{\lambda}_k^{(j)}$, $\lambda_k^{(j)}$ *denote, respectively, the* j*'th element of vectors* $\tilde{\boldsymbol{\lambda}}_k$, $\boldsymbol{\lambda}_k$, *and* σ_1 *is a nonnegative constant.*

Proof: See Appendix. ■

Note that selecting constant sequence $\boldsymbol{b}_k = \boldsymbol{b}$, $k = 1,2,\dots$ trivially satisfies the conditions of Lemma 5. Also, since $\|\cdot\|_2 \le \|\cdot\|_1$ it follows immediately from Lemma 5 that

$$\|\tilde{\boldsymbol{\lambda}}_k - \boldsymbol{\lambda}_k\|_2 \le 2m\alpha\left(\sigma_1/\beta + \sigma_2\right), \quad k = 1,2,\dots \tag{14}$$

This behaviour is illustrated in Figure 3, where it can be seen that the distance between λ_k and $\tilde{\lambda}_k$ remains uniformly bounded over time.

B. Lower Bound for Average Lagrangian

The following lower bound will play a prominent role in later proofs. ****Note: We could move this Lemma to the next section, i.e., after the constrained problem setup.**

Lemma 6 (Lower Bound)**.** *Let* $f : \mathbb{R}^n \to \mathbb{R}$, $g^{(i)} : \mathbb{R}^n \to \mathbb{R}$, $i = 1,\dots,m$ *be convex functions and let* $C \subset \mathbb{R}^n$ *be a convex set. Let Lagrangian* $L(\boldsymbol{z},\boldsymbol{\lambda}) := f(\boldsymbol{z}) + \boldsymbol{\lambda}^T\boldsymbol{g}(\boldsymbol{z})$ *where* $\boldsymbol{g}(\cdot) := [g^{(1)}(\cdot),\dots,g^{(m)}(\cdot)]^T \in \mathbb{R}^m$, $\boldsymbol{\lambda} \in \mathbb{R}^m_+$, *and assume that* $f^* := L(\boldsymbol{z}^*,\boldsymbol{\lambda}^*)$ *is a saddle point,* i.e., $L(\boldsymbol{z}^*,\boldsymbol{\lambda}) \le L(\boldsymbol{z}^*,\boldsymbol{\lambda}^*) \le L(\boldsymbol{z},\boldsymbol{\lambda}^*)$. *Consider update*

$$\boldsymbol{\lambda}_{k+1} = [\boldsymbol{\lambda}_k + \alpha\partial_{\boldsymbol{\lambda}} L(\boldsymbol{z}_{k+1},\boldsymbol{\lambda})]^{[0,\bar{\lambda})} = [\boldsymbol{\lambda}_k + \alpha\boldsymbol{g}(\boldsymbol{z}_{k+1})]^{[0,\bar{\lambda})}$$

[1]Since $\boldsymbol{A}\boldsymbol{z}_{k+1}$ and $\boldsymbol{b}$ can both be negative valued we can think of $\boldsymbol{A}\boldsymbol{z}_{k+1}$ and $\boldsymbol{b}$ as either arrivals or departures, or any combination of arrivals and departures that is convenient – all that matters is the net queue increment $\boldsymbol{A}\boldsymbol{z}_{k+1} - \boldsymbol{b}$.

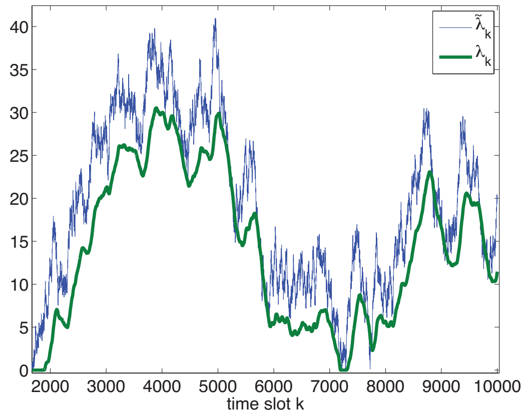


Fig. 3: Example realisations of $\tilde{\lambda}_k$ (thin line) and λ_k (thicker line) for a queue with service rate $b = 0.51$, i.i.d. equiprobable {0,1} arrivals (so mean is 0.5) and parameters $\alpha = 1$, $\beta = 0.01$.

with constant step size $\alpha > 0$ and $\{\boldsymbol{z}_k\}$ is an arbitrary sequence of points from set C. Then,

$$-\frac{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}^*\|_2^2}{2\alpha k} - \frac{\alpha}{2}\sigma_3^2 \le \frac{1}{k}\sum_{i=1}^{k}(\boldsymbol{\lambda}_i - \boldsymbol{\lambda}^*)^T \boldsymbol{g}(\boldsymbol{z}_{i+1}) \quad (15)$$

$$\le \frac{1}{k}\sum_{i=1}^{k} L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - f^* \quad (16)$$

where $\sigma_3 := \max_{\boldsymbol{z}\in C}\|\boldsymbol{g}(\boldsymbol{z})\|_2$.

Proof: We have that

$$\begin{aligned}
&\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^*\|_2^2 \\
&\quad = \|[\boldsymbol{\lambda}_k + \alpha \boldsymbol{g}(\boldsymbol{z}_{k+1})]^{[0,\bar{\lambda})} - \boldsymbol{\lambda}^*\|_2^2 \\
&\quad \le \|\boldsymbol{\lambda}_k + \alpha \boldsymbol{g}(\boldsymbol{z}_{k+1}) - \boldsymbol{\lambda}^*\|_2^2 \\
&\quad \le \|\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^*\|_2^2 + 2\alpha(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}^*)^T \boldsymbol{g}(\boldsymbol{z}_{k+1}) + \alpha^2\sigma_3^2.
\end{aligned}$$

Applying the latter argument recursively for $i = 1, \ldots, k$ yields $\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^*\|_2^2 \le \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}^*\|_2^2 + 2\alpha\sum_{i=1}^{k}(\boldsymbol{\lambda}_i - \boldsymbol{\lambda}^*)^T \boldsymbol{g}(\boldsymbol{z}_{i+1}) + \alpha^2 k\sigma_3^2$. Rearranging terms and using the fact that $\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}^*\|_2^2 \ge 0$ we obtain

$$-\frac{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}^*\|_2^2}{2\alpha} - \frac{\alpha}{2}k\sigma_3^2 \le \sum_{i=1}^{k}(\boldsymbol{\lambda}_i - \boldsymbol{\lambda}^*)^T \boldsymbol{g}(\boldsymbol{z}_{i+1}). \quad (17)$$

To upper bound the RHS observe that

$$\begin{aligned}
&(\boldsymbol{\lambda}_i - \boldsymbol{\lambda}^*)^T \boldsymbol{g}(\boldsymbol{z}_{i+1}) \\
&\quad = (\boldsymbol{\lambda}_i - \boldsymbol{\lambda}^*)^T \boldsymbol{g}(\boldsymbol{z}_{i+1}) + L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) \\
&\quad = L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}^*) \le L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - f^*
\end{aligned}$$

where the last inequality follows from the saddle point property of $L(\boldsymbol{z}^*, \boldsymbol{\lambda}^*)$. It follows that

$$\sum_{i=1}^{k}(\boldsymbol{\lambda}_i - \boldsymbol{\lambda}^*)^T \boldsymbol{g}(\boldsymbol{z}_{i+1}) \le \sum_{i=1}^{k} L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - f^*. \quad (18)$$

Finally, combining (17) and (18) and dividing by k yields the stated result. ■

V. Sequences of Non-Convex Optimisations & Constrained Convex Optimisation

We are now in a position to present our main results. Consider the constrained convex optimisation P:

$$\begin{aligned}
&\underset{\boldsymbol{z}\in C}{\text{minimise}} && f(\boldsymbol{z}) \\
&\text{subject to} && g^{(i)}(\boldsymbol{z}) \le 0, \qquad i = 1, 2, \ldots, m
\end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g^{(i)} : \mathbb{R}^n \to \mathbb{R}$ are convex functions with bounded curvature and, respectively, curvature constants μ_f and $\mu_{g^{(i)}}$, $i = 1, \ldots, m$, $C := \text{conv}(D)$, action set $D \subset \mathbb{R}^n$ consisting of a finite set of points. Let $C_0 := \{\boldsymbol{z} : \boldsymbol{z} \in C, g^{(i)}(\boldsymbol{z}) \le 0, i = 1, 2, \ldots, m\}$ denote the set of feasible points, which we will assume has non-empty relative interior (*i.e.*, a Slater point exists). Let $C^* := \arg\min_{\boldsymbol{z}\in C_0} f(\boldsymbol{z}) \subset C_0$ be the set of optima and $f^* := f(\boldsymbol{z}^*)$, $\boldsymbol{z}^* \in C^*$.

In the next sections we introduce a generalised dual subgradient approach for finding solutions to optimisation P which, as we will see, includes the classical convex dual subgradient method as a special case.

A. Lagrangian Penalty

As in classical convex optimisation we define Lagrangian $L(\boldsymbol{z}, \boldsymbol{\lambda}) := f(\boldsymbol{z}) + \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{z})$ where $\boldsymbol{g}(\cdot) := [g^{(1)}(\cdot), \ldots, g^{(m)}(\cdot)]^T \in \mathbb{R}^m$ and $\boldsymbol{\lambda} \in \mathbb{R}_+^m$. Since set C has non-empty relative interior, the Slater condition is satisfied and strong duality holds. That is, $\min_{\boldsymbol{z}\in C}\max_{\boldsymbol{\lambda}\succeq 0} L(\boldsymbol{z}, \boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}\succeq 0}\min_{\boldsymbol{z}\in C} L(\boldsymbol{z}, \boldsymbol{\lambda}) = f^*$.

1) Lagrangian Bounded Curvature: As already noted, bounded curvature plays a key role in ensuring convergence to an optimum when selecting from a discrete set of actions. For any two points $\boldsymbol{z}, \boldsymbol{z} + \boldsymbol{\delta} \in C$ we have that

$$L(\boldsymbol{z} + \boldsymbol{\delta}, \boldsymbol{\lambda}) \le L(\boldsymbol{z}, \boldsymbol{\lambda}) + \partial_{\boldsymbol{z}} L(\boldsymbol{z}, \boldsymbol{\lambda})^T \boldsymbol{\delta} + \mu_L\|\boldsymbol{\delta}\|_2^2, \quad (19)$$

where

$$\mu_L = \mu_f + \boldsymbol{\lambda}^T \boldsymbol{\mu}_g$$

with $\boldsymbol{\mu}_g := [\mu_{g^{(1)}}, \ldots, \mu_{g^{(m)}}]^T$. It can be seen that the curvature constant μ_L of the Lagrangian depends on the multiplier $\boldsymbol{\lambda}$. Since set $\boldsymbol{\lambda} \succeq 0$ is unbounded, it follows that the Lagrangian does not have bounded curvature on this set unless $\boldsymbol{\mu}_g = \mathbf{0}$ (corresponding to the special case where the constraints are linear).

Fortunately, when set C has non-empty relative interior it is not necessary to consider the full set $\boldsymbol{\lambda} \succeq 0$ and a bounded subset is sufficient. The following result corresponds to [21, Lemma 1], and we include a proof for completeness.

Lemma 7 (Bounded Multipliers)**.** *Let $q(\boldsymbol{\lambda}) := \min_{\boldsymbol{z}\in C} L(\boldsymbol{z}, \boldsymbol{\lambda})$ and $\boldsymbol{\lambda}^* \in \arg\max_{\boldsymbol{\lambda}\succeq 0} q(\boldsymbol{\lambda})$. Suppose set C has non-empty relative interior. Then*

$$\|\boldsymbol{\lambda}^*\|_2 \le \frac{f(\bar{\boldsymbol{z}}) - q(\boldsymbol{\lambda})}{\xi} \quad (20)$$

where $\xi = -\min_{1\le i\le m} g^{(i)}(\bar{\boldsymbol{z}})$, $\bar{\boldsymbol{z}} \in \text{relint}(C)$.

Proof: Since $\bar{\boldsymbol{z}} \in \text{relint}(C)$ we have that $g^{(i)}(\bar{\boldsymbol{z}}) < 0$ for all $i = 1, \ldots, m$. Now, $q(\boldsymbol{\lambda}^*) \le L(\bar{\boldsymbol{z}}, \boldsymbol{\lambda}^*) = f(\bar{\boldsymbol{z}}) +$

$\sum_{i=1}^m \lambda^{*(i)} g^{(i)}(\bar{z})$ and therefore $-\sum_{i=1}^m \lambda^{*(i)} g^{(i)}(\bar{z}) \le f(\bar{z}) - q(\boldsymbol{\lambda}^*)$. Next, define $\xi = -\min_{1\le i\le m} g^{(i)}(\bar{z})$ and see $\xi \sum_{i=1}^m \lambda^{*(i)} \le f(\bar{z}) - q(\boldsymbol{\lambda}^*)$. Finally, since $-q(\boldsymbol{\lambda}^*) \le -q(\boldsymbol{\lambda})$ and $\|\cdot\|_2 \le \|\cdot\|_1$ the stated result now follows. ■

From Lemma 7 it follows that there exists a constant $\bar{\lambda} \in [0,\infty)$ such that $\boldsymbol{\lambda}^* \preceq \bar{\lambda}\mathbf{1} := \bar{\boldsymbol{\lambda}}$. That is, it is sufficient to confine consideration to the Lagrangian on bounded set $0 \preceq \boldsymbol{\lambda} \preceq \bar{\boldsymbol{\lambda}}$ since $\max_{0\preceq\boldsymbol{\lambda}\preceq\bar{\boldsymbol{\lambda}}} \min_{\boldsymbol{z}\in C} L(\boldsymbol{z},\boldsymbol{\lambda}) = f^*$. On this set the Lagrangian has bounded curvature with curvature constant

$$\bar{\mu}_L = \mu_f + \bar{\boldsymbol{\lambda}}^T \boldsymbol{\mu}_g.$$

2) Non-Convex Dual Subgradient Update: Now consider the following sequence of non-convex optimisations $\{\hat{P}_k\}$:

$$\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} \; L((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}, \boldsymbol{\lambda}_k) \tag{21}$$

$$\boldsymbol{z}_{k+1} = (1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}_k \tag{22}$$

$$\boldsymbol{\lambda}_{k+1} = [\boldsymbol{\lambda}_k + \alpha \boldsymbol{g}(\boldsymbol{z}_{k+1})]^{[0,\bar{\lambda})} \tag{23}$$

where $\beta \in (0,1)$, $\alpha > 0$. Update (23) ensures $0 \preceq \boldsymbol{\lambda}_{k+1} \preceq \bar{\boldsymbol{\lambda}}$.

Theorem 3. *Consider convex optimisation P and the associated sequence of non-convex optimisations $\{\hat{P}_k\}$. Suppose parameters β and α are selected sufficiently small that*

$$0 < \beta \le (1-\gamma)\min\{\epsilon/(\bar{\mu}_L \bar{x}_D^2), 1\}, \tag{24}$$

$$0 < \alpha \le \gamma_1\gamma\beta\epsilon/\sigma_3^2 \tag{25}$$

with $\epsilon > 0$, $0 < \gamma < 1$, $0 < \gamma_1 < 1/2$, $\bar{\mu}_L = \mu_f + \bar{\boldsymbol{\lambda}}^T \boldsymbol{\mu}_g$ and $\bar{\boldsymbol{\lambda}} \succeq \boldsymbol{\lambda}^*$. Then, there exists a $\bar{k}$ such that for all $k \ge \bar{k}$ the sequence of solutions $\{\boldsymbol{z}_k\}$ to sequence of optimisations $\{\hat{P}_k\}$ satisfies:

$$-\frac{\|\boldsymbol{\lambda}_{\bar{k}} - \boldsymbol{\lambda}^*\|_2^2}{2\alpha k} - \frac{\alpha}{2}\sigma_3^2 \le \frac{1}{k}\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - f^* \le 2\epsilon \tag{26}$$

where $\sigma_3 := \max_{\boldsymbol{z}\in C} \|\boldsymbol{g}(\boldsymbol{z})\|_2$.

Proof: $L(\cdot, \boldsymbol{\lambda}_k)$ is convex and has uniformly bounded curvature on C with curvature constant $\bar{\mu}_L = \mu_f + \bar{\boldsymbol{\lambda}}^T \boldsymbol{\mu}_g$. Further, $\|L(\boldsymbol{z},\boldsymbol{\lambda}_{k+1}) - L(\boldsymbol{z},\boldsymbol{\lambda}_k)\|_2 = \|(\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k)^T \boldsymbol{g}(\boldsymbol{z})\|_2 \le \|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\|_2 \sigma_3 \le \alpha\sigma_3^2 \le \gamma_1\gamma\beta\epsilon$. Hence, by Theorem 1 for k sufficiently large ($k \ge \bar{k}$) we have that $L(\boldsymbol{z}_{k+1},\boldsymbol{\lambda}_k) - L(\boldsymbol{y}_k^*,\boldsymbol{\lambda}_k) \le 2\epsilon$, where $\boldsymbol{y}_k^* \in \arg\min_{\boldsymbol{z}\in C} L(\boldsymbol{z},\boldsymbol{\lambda}_k)$. By the saddle point property $L(\boldsymbol{z}^*,\boldsymbol{\lambda}) \le L(\boldsymbol{z}^*,\boldsymbol{\lambda}^*) \le L(\boldsymbol{z},\boldsymbol{\lambda}^*)$ and the fact that $f^* = L(\boldsymbol{z}^*,\boldsymbol{\lambda}^*)$, right-hand inequality in (26) now follows. The left-hand inequality in (26) follows by application of Lemma 6. ■

3) Generalised Update: Let $C' \subseteq \mathrm{conv}(D)$ be any subset of the convex hull of action set D, including the empty set. Since $\min_{\boldsymbol{x}\in C'\cup D} \; L((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}, \boldsymbol{\lambda}_k) \le \min_{\boldsymbol{x}\in D} \; L((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}, \boldsymbol{\lambda}_k)$, we can immediately generalise update (21) to

$$\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in C'\cup D} L((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}, \boldsymbol{\lambda}_k) \tag{27}$$

and Theorem 3 will continue to apply. Selecting C' equal to the empty set we recover (21) as a special case. Selecting $C' = \mathrm{conv}(D)$ we recover the classical convex dual subgradient update as a special case. Update (27) therefore naturally generalises both the classical convex dual subgradient update and non-convex update (21).

B. Alternative Update

Note that, by replacing use of Theorem 1 by Theorem 2 in the proof, we can replace update (21) by its non-convex Frank-Wolfe alternative,

$$\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} \; \partial_z L(\boldsymbol{z}_k, \boldsymbol{\lambda}_k)^T \boldsymbol{x} \tag{28}$$

$$= \arg\min_{\boldsymbol{x}\in D} \; (\partial f(\boldsymbol{z}_k) + \boldsymbol{\lambda}_k^T \partial \boldsymbol{g}(\boldsymbol{z}_k))^T \boldsymbol{x}. \tag{29}$$

That is, we have:

Corollary 2. *Consider the setup in Theorem 3 but with update (21) replaced by (29). Then,*

$$-\frac{\|\boldsymbol{\lambda}_{\bar{k}} - \boldsymbol{\lambda}^*\|_2^2}{2\alpha k} - \frac{\alpha}{2}\sigma_3^2 \le \frac{1}{k}\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - f^* \le 2\epsilon.$$

C. Discrete Multiplier Update & Queues

In (27) set $C' \cup D$ extends the primal optimisation step in the classical dual subgradient update to encompass discrete as well as continuous actions. The multiplier update (23) continues, however, to use continuous-valued average action quantity $\boldsymbol{z}_{k+1}$. Using Lemma 5 we can also extend the multiplier update to encompass discrete as well as continuous actions.

1) Linear Constraints: Consider the special case of optimisation P where the constraints are linear. That is, $g^{(i)}(\boldsymbol{z}) = (\boldsymbol{a}^{(i)})^T \boldsymbol{z} - b^{(i)}$ where $\boldsymbol{a}^{(i)} \in \mathbb{R}^n$ and $b^{(i)} \in \mathbb{R}$, $i = 1, \ldots, m$. Gathering vectors $(\boldsymbol{a}^{(i)})^T$ together as the rows of matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and collecting additive terms $b^{(i)}$ into vector $\boldsymbol{b} \in \mathbb{R}^m$, the linear constraints can then be written as $\boldsymbol{A}\boldsymbol{z} \preceq \boldsymbol{b}$. Consider the following sequence of non-convex optimisations $\{\tilde{P}_k^L\}$:

$$\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} \; L((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}, \tilde{\boldsymbol{\lambda}}_k) \tag{30}$$

$$= \arg\min_{\boldsymbol{x}\in D} f((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}) + \beta\tilde{\boldsymbol{\lambda}}_k^T \boldsymbol{A}\boldsymbol{x} \tag{31}$$

$$\boldsymbol{z}_{k+1} = (1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}_k \tag{32}$$

$$\tilde{\boldsymbol{\lambda}}_{k+1} = [\tilde{\boldsymbol{\lambda}}_k + \alpha(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}_k)]^+ \tag{33}$$

where $\beta \in (0,1)$, $\alpha > 0$ and $\{\boldsymbol{b}_k\}$ is a sequence of points from set $E \subseteq \mathbb{R}^m$ such that $\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^k \boldsymbol{b}_i = \boldsymbol{b}$. Update (31) is obtained from (30) by retaining only the parts of $L((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}, \tilde{\boldsymbol{\lambda}}_k)$ which depend on $\boldsymbol{x}$ *i.e.*, dropping constant terms which do not change the solution to the optimisation. Note that (31) does not involve $\boldsymbol{b}$ or $\boldsymbol{b}_k$.

Observe that the multiplier update (33) is now modified to make use of the discrete action $\boldsymbol{x}_k$ and to allow a time-varying, possibly discrete, additive term $\boldsymbol{b}_k$. We have also removed the $\bar{\lambda}$ upper limit since with linear constraints the Lagrangian has bounded curvature even when the multipliers may be unbounded. Provided α/β is sufficiently small and $\frac{1}{k}\sum_{i=1}^k \boldsymbol{b}_i$ approaches limit $\boldsymbol{b}$ sufficiently fast, by Lemma 5 we know that the discrete-valued multiplier $\tilde{\boldsymbol{\lambda}}_k$ generated by (33) will remain close in value to the continuously valued multiplier $\boldsymbol{\lambda}_k$ generated by (23). It is unsurprising therefore that sequence $\{\tilde{P}_k^L\}$ will also converge to the solution of optimisation P:

Theorem 4. *Consider optimisation P and suppose that the constraints are linear, $\boldsymbol{A}\boldsymbol{z} \preceq \boldsymbol{b}$. Consider also the associated sequence of non-convex optimisations $\{\tilde{P}_k^L\}$. Suppose parameters β and α are selected sufficiently small that $\beta \leq (1-\gamma)\min\{\epsilon/(\mu_L \bar{x}_D^2), 1\}$, $0 < \gamma < 1$, $\epsilon > 0$ and $\alpha \leq \gamma_1\gamma\beta\epsilon/\sigma_3^2$, $0 \leq \gamma_1 < 1/2$. Also that sequence $\{\boldsymbol{b}_k\}$ of points from set $E \subseteq \mathbb{R}^m$ satisfies $\lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^k \boldsymbol{b}_i = \boldsymbol{b}$ and $|(\frac{1}{k}\sum_{i=1}^k b_i^{(j)}) - b^{(j)}| \leq \sigma_2/k$, $j = 1,\dots,m$. Then there exists $\bar{k}$ such that for all $k \geq \bar{k}$ the sequence of solutions $\{\boldsymbol{z}_k\}$ to sequence of optimisations $\{\tilde{P}_k^L\}$ satisfies:*

$$
\begin{aligned}
-\frac{\|\tilde{\boldsymbol{\lambda}}_{\bar{k}} - \boldsymbol{\lambda}^*\|_2^2}{2\alpha k} - \frac{\alpha}{2}\sigma_3^2 - 2m\alpha\left(\sigma_1/\beta + \sigma_2\right)\sigma_3 \\
\leq \frac{1}{k}\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \tilde{\boldsymbol{\lambda}}_i) - f^* \leq 2\epsilon. \qquad (34)
\end{aligned}
$$

Proof: Define auxiliary update $\boldsymbol{\lambda}_{k+1} = [\boldsymbol{\lambda}_k + \alpha(\boldsymbol{A}\boldsymbol{z}_{k+1} - \boldsymbol{b})]^+$ where sequence $\{\boldsymbol{z}_k\}$ is from optimisations $\{\tilde{P}_k^L\}$ and with initial condition $\boldsymbol{\lambda}_1 = \tilde{\boldsymbol{\lambda}}_1$. The $\boldsymbol{\lambda}_k$ in this update is not used in optimisations $\{\tilde{P}_k^L\}$, but will nevertheless prove useful. Recall $L(\boldsymbol{z}_{k+1}, \tilde{\boldsymbol{\lambda}}_k) = f(\boldsymbol{z}_{k+1}) + \tilde{\boldsymbol{\lambda}}_k^T(\boldsymbol{A}\boldsymbol{z}_{k+1} - \boldsymbol{b})$. Now $\|L(\boldsymbol{z}, \tilde{\boldsymbol{\lambda}}_{k+1}) - L(\boldsymbol{z}, \tilde{\boldsymbol{\lambda}}_k)\|_2 \leq \|\tilde{\boldsymbol{\lambda}}_{k+1} - \tilde{\boldsymbol{\lambda}}_k\|_2\sigma_3 \leq \alpha\sigma_3^2 \leq \gamma_1\gamma\beta\epsilon$ and so by Theorem 1, for $i \geq \bar{k}$,

$$
\frac{1}{k}\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \tilde{\boldsymbol{\lambda}}_i) - f^* \leq 2\epsilon. \qquad (35)
$$

Since Lemma 6 holds for any interval $i = \bar{k},\dots,\bar{k}+k$ we have

$$
-\frac{\|\boldsymbol{\lambda}_{\bar{k}} - \boldsymbol{\lambda}^*\|_2^2}{2\alpha k} - \frac{\alpha}{2}\sigma_3^2 \leq \frac{1}{k}\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - f^* \qquad (36)
$$

where σ_3 is finite since constraints are linear, C is bounded and $|\sum_{i=1}^k b_i^{(j)} - b^{(j)}| \leq \sigma_2$. Now, see that

$$
\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - L(\boldsymbol{z}_{i+1}, \tilde{\boldsymbol{\lambda}}_i) = \sum_{i=\bar{k}}^{k+\bar{k}} (\boldsymbol{\lambda}_i - \tilde{\boldsymbol{\lambda}}_i)^T(\boldsymbol{A}\boldsymbol{z}_{i+1} - \boldsymbol{b}).
$$

Using Lemma 5 and the Cauchy-Schwarz inequality we have that

$$
\begin{aligned}
\sum_{i=\bar{k}}^{k+\bar{k}} (\tilde{\boldsymbol{\lambda}}_i - \boldsymbol{\lambda}_i)^T(\boldsymbol{A}\boldsymbol{z}_{i+1} - \boldsymbol{b}) &\geq -\sum_{i=\bar{k}}^{k+\bar{k}} \|\tilde{\boldsymbol{\lambda}}_i - \boldsymbol{\lambda}_i\|_2 \|\boldsymbol{A}\boldsymbol{z}_{i+1} - \boldsymbol{b}\|_2 \\
&\geq -2m\alpha\left(\sigma_1/\beta + \sigma_2\right)k\sigma_3. \qquad (37)
\end{aligned}
$$

Dividing (37) by k and combining it with (35) and (36) yields the desired result. ■

Observe that in Theorem 4 the LHS is looser than in Theorem 3 as it involves the additional term $2m\alpha(\sigma_1/\beta + \sigma_2)\sigma_3$. That is, a smaller value of α may be needed to achieve accuracy with a discrete multiplier update than when a continuous valued multiplier update is used.

2) Queues: As already noted in Section IV-A, $\tilde{\boldsymbol{\lambda}}_k$ can be thought of as the α scaled occupancy of a set of queues with net increment $\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}_k$. That is, use of discrete update (33) induces an equivalence between multipliers and the occupancy of an associated set of queues. Note that this observation complements [13][Theorem 1], which states that a discrete queue update tends on average to drift towards the optimal multiplier value. As we will see later, in network flow problems these queues can be identified with physical link queues. However, since the formulation of network flow constraints is not unique, and changing how these constraints are formulated changes the multiplier update, some care may be necessary to select the constraints in a way that is congruent with the physical queues in a particular network. Further, since the multiplier queue occupancy is given by $\tilde{\boldsymbol{Q}}_k = \tilde{\boldsymbol{\lambda}}_k/\alpha$, for constraints where the multipler is non-zero at the optimum the associated queue occupancy will necessarily grow as subgradient step size α is decreased. Since a small step size is generally needed in order to converge to a small ball around the optimum, this indicates that a fundamental trade-off may exist between queue occupancy and optimality.

3) Alternative Update: Once again, by replacing use of Theorem 1 by Theorem 2 in the proof, we can replace update (31) by its non-convex Frank-Wolfe alternative,

$$
\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} \partial f(\boldsymbol{z}_k)^T\boldsymbol{x} + \tilde{\boldsymbol{\lambda}}_k^T\boldsymbol{A}\boldsymbol{x}. \qquad (38)
$$

That is, we have:

Corollary 3. *Consider the setup in Theorem 4 but with update (31) replaced by (38). Then,*

$$
\begin{aligned}
-\frac{\|\tilde{\boldsymbol{\lambda}}_{\bar{k}} - \boldsymbol{\lambda}^*\|_2^2}{2\alpha k} - \frac{\alpha}{2}\sigma_3^2 - 2m\alpha\left(\sigma_1/\beta + \sigma_2\right)\sigma_3 \\
\leq \frac{1}{k}\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \tilde{\boldsymbol{\lambda}}_i) - f^* \leq 2\epsilon. \qquad (39)
\end{aligned}
$$

4) Nonlinear Constraints: The foregoing is for linear constraints. When the constraints are nonlinear, since $g^{(i)}(\cdot)$ has bounded curvature,

$$
\begin{aligned}
&g^{(i)}(\boldsymbol{z}_{i+1}) - g^{(i)}(\boldsymbol{x}_i) \\
&\quad \leq \partial g^{(i)}(\boldsymbol{x}_i)^T(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i) + \mu_{g^{(i)}}\|\boldsymbol{z}_{i+1} - \boldsymbol{x}_i\|_2^2 \\
&\quad \leq \nu_{g^{(i)}}\|\boldsymbol{z}_{i+1} - \boldsymbol{x}_i\|_2 + \mu_{g^{(i)}}\beta^2\bar{x}_D^2
\end{aligned}
$$

where $\nu_{g^{(i)}}$ is the Lipschitz constant of $g^{(i)}(\cdot)$ on C (which exists by Lemma 1). It follows that for all $k \geq 1$

$$
\begin{aligned}
&\left\|\sum_{i=1}^k g^{(i)}(\boldsymbol{z}_{i+1}) - g^{(i)}(\boldsymbol{x}_i)\right\|_2 \\
&\quad \leq \nu_{g^{(i)}}\left\|\sum_{i=1}^k (\boldsymbol{z}_{i+1} - \boldsymbol{x}_i)\right\|_2 + k\mu_{g^{(i)}}\beta^2\bar{x}_D^2.
\end{aligned}
$$

For the discrete and continuous multiplier updates to remain close, we require the RHS to be uniformly bounded for all k. This requires $k\mu_{g^{(i)}}\beta^2\bar{x}_D^2$ to be uniformly bounded, which might be achieved for example by selecting β to be diminishing as $1/\sqrt{k}$ rather than as a constant. However, while Theorem 1 can be readily extended to diminishing β, the extension of the other basic building block in our analysis, Lemma 6, to allow diminishing β is not straightforward. We therefore leave this extension to future work.

VI. Stochastic Optimisation

The analysis in the preceeding Section V is for deterministic optimisation problems. However, it can be readily extended to a class of stochastic optimisations that encompasses those considered in max-weight approaches.

A. Stochastic Contraints

1) Linear Constraints: Of particular interest, in view of the equivalence which has been established between multiplier updates and queues, is accommodating stochastic queue arrivals. When the constraints are linear $\boldsymbol{A}\boldsymbol{z} \preceq \boldsymbol{b}$, and $\{\boldsymbol{b}_k\}$ is a sequence such that $\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{b}_i$ converges sufficiently quickly to $\boldsymbol{b}$, then by Theorem 4 the sequence of non-convex optimisations $\{\tilde{P}_k^L\}$ converges to the solution of optimisation problem P. Since this holds for all admissible sequences $\{\boldsymbol{b}_k\}$, this of course includes sample paths of a stochastic process.

Let $\{\boldsymbol{B}_k\}$ be a stochastic process with realisations of $\boldsymbol{B}_k$ taking values in $E \subseteq \mathbb{R}^m$ and with mean $\boldsymbol{b} \in \mathbb{R}^m$. Let $p_k := \mathrm{Prob}(\|(\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{B}_i) - \boldsymbol{b}\|_k \le \sigma_2/k)$. Note that, by central limit arguments, for many stochastic processes $\lim_{k\to\infty} p_k = 0$. Let $\{\boldsymbol{b}_i\}_{i=1}^k$ denote a realisation of length k and E^k the set of possible realisations of length k. Fraction p_k of these realisations satisfy $\|(\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{b}_i) - \boldsymbol{b}\|_k \le \sigma_2/k$, *i.e.,* fraction p_k of realisations satisfy the conditions of Theorem 4. We therefore have the following corollary to Theorem 4.

Corollary 4. *Consider the setup in Theorem 4. Suppose that sequence $\{\boldsymbol{b}_k\}$ is a realisation of a stochastic process $\{\boldsymbol{B}_k\}$ with mean $\boldsymbol{b} \in \mathbb{R}^m$. Let $p_k := \mathrm{Prob}(\|(\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{B}_i) - \boldsymbol{b}\|_\infty \le \sigma_2/k)$. Then there exists $\bar{k}$ such that with probability p_k for all $k \ge \bar{k}$ the sequence of solutions $\{\boldsymbol{z}_k\}$ to sequence of optimisations $\{\tilde{P}_k^L\}$ satisfies (34).*

Recall that set E need not be convex and may consist of a finite number of points. For example, we might have $E = \{0,1\}^m$ and think of $\boldsymbol{B}_k$ as random packet arrivals/departures at time step k and $\boldsymbol{b}$ as the mean packet arrival/departure rate. Note that there is no requirement for stochastic process $\{\boldsymbol{B}_k\}$ to be i.i.d. or for any of its properties, other than that feasible set $\boldsymbol{A}\boldsymbol{z} \preceq \boldsymbol{b}$ has non-empty relative interior, to be known in advance in order to construct solution sequence $\{\tilde{P}_k^L\}$. Note also that while we require an interior (Slater) point to exist for $\boldsymbol{A}\boldsymbol{z} \preceq \boldsymbol{b}$ we do not require this to be the case for constraint $\boldsymbol{A}\boldsymbol{z} \preceq \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{B}_i$ for finite k.

2) Nonlinear constraints: We can partially extend the above analysis to nonlinear constraints of the form $\boldsymbol{g}(\boldsymbol{z}) \preceq \boldsymbol{b}$ (partial in the sense that we retain use of averaged quantity $\boldsymbol{z}_k$ in the multiplier update, but allow $\boldsymbol{b}$ to be replaced by a sequence). Namely, consider the sequence of non-convex optimisations $\{\hat{P}_k^{NL}\}$:

$$\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} L((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}, \tilde{\boldsymbol{\lambda}}_k) \tag{40}$$

$$= \arg\min_{\boldsymbol{x}\in D} f((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}) + \tilde{\boldsymbol{\lambda}}_k^T \boldsymbol{g}((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x})$$

$$\boldsymbol{z}_{k+1} = (1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}_k \tag{41}$$

$$\tilde{\boldsymbol{\lambda}}_{k+1} = [\tilde{\boldsymbol{\lambda}}_k + \alpha(\boldsymbol{g}(\boldsymbol{z}_{k+1}) - \boldsymbol{b}_k)]^{[0,\bar{\lambda})}. \tag{42}$$

Theorem 5. *Consider optimisation P with modified constraints $\boldsymbol{g}(\boldsymbol{z}) \preceq \boldsymbol{b}$, and consider associated sequence of non-convex optimisations $\{\hat{P}_k^{NL}\}$. Suppose parameters β and α satisfy the conditions in Theorem 3. Also that sequence $\{\boldsymbol{b}_k\}$ of points from set $E \subseteq \mathbb{R}^m$ satisfies $\lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{b}_i = \boldsymbol{b}$ and $|(\frac{1}{k}\sum_{i=1}^{k}b_i^{(j)}) - b^{(j)}| \le \sigma_2/k$, $j = 1,\ldots,m$. Then there exists $\bar{k}$ such that for all $k \ge \bar{k}$ the sequence of solutions $\{\boldsymbol{z}_k\}$ to sequence of optimisations $\{\tilde{P}_k^{NL}\}$ satisfies:*

$$-\frac{\|\tilde{\boldsymbol{\lambda}}_{\bar{k}} - \boldsymbol{\lambda}^*\|_2^2}{2\alpha k} - 2\alpha\sigma_3\left(\frac{\sigma_3}{2} + m\sigma_2\right) \le \frac{1}{k}\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \tilde{\boldsymbol{\lambda}}_i) - f^* \le 2\epsilon. \tag{43}$$

Proof: The proof closely parallels that of Theorem 4, but due to the nonlinearity of $\boldsymbol{g}(\cdot)$ can only make use of Lemma 4 rather than the stronger Lemma 5. Define auxiliary update $\boldsymbol{\lambda}_{k+1} = [\boldsymbol{\lambda}_k + \min\{\alpha(\boldsymbol{g}(\boldsymbol{z}_{k+1}) - \boldsymbol{b}), \bar{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_k\}]^{[0,\bar{\lambda})}$ where sequence $\{\boldsymbol{z}_k\}$ is from optimisations $\{\tilde{P}_k^{NL}\}$ and with initial condition $\boldsymbol{\lambda}_1 = \tilde{\boldsymbol{\lambda}}_1$. Since $\|L(\boldsymbol{z}, \tilde{\boldsymbol{\lambda}}_{k+1}) - L(\boldsymbol{z}, \tilde{\boldsymbol{\lambda}}_k)\|_2 \le \|\tilde{\boldsymbol{\lambda}}_{k+1} - \tilde{\boldsymbol{\lambda}}_k\|_2\sigma_3 \le \alpha\sigma_3^2 \le \gamma_1\gamma\beta\epsilon$, by Theorem 1, $\frac{1}{k}\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \tilde{\boldsymbol{\lambda}}_i) - f^* \le 2\epsilon$ for $i \ge \bar{k}$. Also, by Lemma 6, $-\frac{\|\boldsymbol{\lambda}_{\bar{k}} - \boldsymbol{\lambda}^*\|_2^2}{2\alpha k} - \frac{\alpha}{2}\sigma_3^2 \le \frac{1}{k}\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - f^*$. Now, $\sum_{i=\bar{k}}^{k+\bar{k}} L(\boldsymbol{z}_{i+1}, \boldsymbol{\lambda}_i) - L(\boldsymbol{z}_{i+1}, \tilde{\boldsymbol{\lambda}}_i) = \sum_{i=\bar{k}}^{k+\bar{k}}(\boldsymbol{\lambda}_i - \tilde{\boldsymbol{\lambda}}_i)^T(\boldsymbol{g}(\boldsymbol{z}_{i+1}) - \boldsymbol{b}) \ge -\sigma_3\sum_{i=\bar{k}}^{k+\bar{k}}\|\tilde{\boldsymbol{\lambda}}_i - \boldsymbol{\lambda}_i\|_2 \ge -2m\alpha\sigma_2 k\sigma_3$ by Lemma 4. Combining the above inequalities yields the result. ■

Corollary 5. *Consider the setup in Theorem 5. Suppose that sequence $\{\boldsymbol{b}_k\}$ is a realisation of a stochastic process $\{\boldsymbol{B}_k\}$ with mean $\boldsymbol{b} \in \mathbb{R}^m$. Let $p_k := \mathrm{Prob}(\|(\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{B}_i) - \boldsymbol{b}\|_\infty \le \sigma_2/k)$. Then there exists $\bar{k}$ such that with probability p_k for all $k \ge \bar{k}$ the sequence of solutions $\{\boldsymbol{z}_k\}$ to sequence of optimisations $\{\tilde{P}_k^{NL}\}$ satisfies (43).*

B. Stochastic Actions

Suppose that when at time k we select action $\boldsymbol{x}_k \in D$, the action actually applied is a realisation of random variable $\boldsymbol{Y}_k$ that also takes values in D; this is for simplicity, the extension to random action sets different from D is straightforward). For example, we may select $x_k = 1$ (which might correspond to transmitting a packet) but with some probability actually apply $y_k = 0$ (which might correspond to a transmisison failure/packet loss). Let $p_{\boldsymbol{x}\boldsymbol{y}} := \mathrm{Prob}(\boldsymbol{Y}_k = \boldsymbol{y}|\boldsymbol{x}_k = \boldsymbol{x})$, $\boldsymbol{x}$, $\boldsymbol{y} \in D$ and we assume that this probability distribution is time-invariant *i.e.,* does not depend on k; again, this can be relaxed in the obvious manner.

Assume that the probabilities $p_{\boldsymbol{x}\boldsymbol{y}}$, $\boldsymbol{x}$, $\boldsymbol{y} \in D$ are known. Then $\bar{\boldsymbol{y}}(\boldsymbol{x}) := \mathbb{E}[\boldsymbol{Y}_k|\boldsymbol{x}_k = \boldsymbol{x}] = \sum_{\boldsymbol{y}\in D}\boldsymbol{y}p_{\boldsymbol{x}\boldsymbol{y}}$ can be calculated. The above analysis now carries over unchanged provided we modify the non-convex optimisation from $\min_{\boldsymbol{x}\in D} L((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}, \boldsymbol{\lambda}_k)$ to $\min_{\boldsymbol{x}\in D} L((1-\beta)\boldsymbol{z}_k + \beta\bar{\boldsymbol{y}}(\boldsymbol{x}), \boldsymbol{\lambda}_k)$ and everywhere replace $\boldsymbol{x}_k$ by $\bar{\boldsymbol{y}}(\boldsymbol{x}_k)$. That is, we simply change variables to $\bar{\boldsymbol{y}}$. Note that this relies upon the mapping from $\boldsymbol{x}$ to $\bar{\boldsymbol{y}}$ being known. If this is not the case, then we are entering the realm of stochastic decision problems and we leave this to future work.

VII. Max-Weight Revisited

A. Discussion

Recall the formulation of a queueing network in Section I, where matrix $\boldsymbol{A}$ defines the queue interconnection, with i'th row having a -1 at the i'th entry, 1 at entries corresponding to queues from which packets are sent to queue i, and 0 entries elsewhere. Hence, the queue occupancy evolves as $\boldsymbol{Q}_{k+1} = [\boldsymbol{Q}_k + \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{b}_k]^+$. By Corollary 3 we know that update $\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} \partial f(\boldsymbol{z}_k)^T\boldsymbol{x} + \alpha \boldsymbol{Q}_k^T\boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{z}_{k+1} = (1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}_k$ leads to $\boldsymbol{z}_k$ converging to a ball around the solution to the following convex optimisation,

$$\begin{aligned} \underset{\boldsymbol{z}\in C}{\text{minimise}} \quad & f(\boldsymbol{z}) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b} \preceq 0 \end{aligned}$$

where $C = \text{conv}(D)$, $\{\boldsymbol{b}_k\}$ is any sequence such that $\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^k \boldsymbol{b}_i = \boldsymbol{b}$ and $|(\frac{1}{k}\sum_{i=1}^k b_i^{(j)}) - b^{(j)}| \le \sigma_2/k$, $j = 1,\dots,m$ for some finite $\sigma_2 > 0$.

Observe that this update is identical to the greedy primal-dual max-weight schedule once we identify utility function $U(\cdot)$ with $-f(\cdot)$. However, we have arrived at this from a purely convex optimisation viewpoint and by elementary arguments, without recourse to more sophisticated Lyapunov drift, stochastic queueing theory *etc*. Further, Corollary 3 immediately generalises the max-weight analysis to allow arbitary linear constraints rather than just the specific constraints associated with a queueing network, and beyond this to convex nonlinear constraints with bounded curvature.

In our analysis, the key role played by bounded curvature in non-convex descent is brought to the fore. This property is of course present in existing max-weight results, in the form of a requirement for continuous differentiability of the utility function, but insight into the fundamental nature of this requirement had been lacking. One immediate benefit is the resulting observation that any non-convex update with suitable descent properties can be used, and strong connections are established with the wealth of convex descent methods. For example, by Theorem 4 we can replace update $\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D}(\partial f(\boldsymbol{z}_k) + \boldsymbol{A}\boldsymbol{Q}_k)^T\boldsymbol{x}$ (which is now seen to be a variant of the classical Frank-Wolfe update) with the direct Lagrangian update $\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in D} f(\boldsymbol{z}_k + \beta(\boldsymbol{x} - \boldsymbol{z}_k)) + \beta\boldsymbol{Q}_k^T\boldsymbol{A}\boldsymbol{x}$ to obtain a new class of non-convex algorithms.

B. Example: Privacy-Enhancing Online Scheduling

We present a simple example with a nonlinear constraint. Suppose packets arrive at a queue with inter-arrival times $\{b_k\}$, $k = 1, 2, \dots$ and the queue service is stochastic and selected such that the entropy of the inter-packet times of the outgoing packet stream is at least E in order to provide a degree of resistence to traffic analysis. Note that dummy packets are transmitted when no useful packets are available, so as prevent large inter-arrivals times from propagating to the outgoing packet stream. Time is slotted so b_k, $x_k \in \mathbb{N}$. The packet arrival process is not known in advance, other than the facts that it can be feasibily served, the inter-arrival times have finite mean $\lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^k b_i = b$ and $|(\frac{1}{k}\sum_{i=1}^k b_i) - b| \le \sigma_2/k$ for some finite $\sigma_2 > 0$.

Let set $\{0, 1, \dots, T\}$ denotes the set of inter-packet times that can potentially be scheduled on the outgoing link and consider the following feasibility problem (couched in convex optimisation form),

$$\min_{\boldsymbol{p}\in C} 1 \text{ s.t. } \sum_{i=0}^{T_{max}} p^{(i)} \log p^{(i)} \le -E, \ \sum_{i=0}^{T_{max}} ip^{(i)} \le b$$

where $p^{(i)} = \text{Prob}(x_k = i)$, $C := \{\boldsymbol{p} \in [\delta, 1-\delta]^T, \delta \in (0,1) : \sum_{i=1}^T p^{(i)} \le 1\}$. Notice that we have defined set C so that the first constraint has bounded curvature. If the arrival process $\{b_k\}$ were known in advance, we could solve this optimisation to determine a feasible $\boldsymbol{p}$. When the arrivals are not known in advance, using generalised update (27) by a corollary of Theorem 3 we can instead use the following online update to determine a sequence $\{\boldsymbol{p}_k\}$ that converges to a feasible point.

$$\begin{aligned} &\boldsymbol{x}_k \in \arg\min_{\boldsymbol{p}\in C} \lambda_k \sum_{i=0}^{T_{max}} p^{(i)} \log p^{(i)} + \theta_k \sum_{i=0}^{T_{max}} ip^{(i)} \\ &\boldsymbol{p}_{k+1} = (1-\beta)\boldsymbol{p}_k + \beta\boldsymbol{x}_k \\ &\lambda_{k+1} = \left[\lambda_k + \alpha\left(\sum_{i=0}^{T_{max}} p^{(i)} \log p^{(i)} + E\right)\right]^{[0,\bar{\lambda}]} \\ &\theta_{k+1} = \left[\theta_k + \alpha\left(\sum_{i=0}^{T_{max}} ip^{(i)} - b_k\right)\right]^{[0,\bar{\theta}]}. \end{aligned}$$

By introducing an objective that measures mean delay, we could further extend this setup to minimise delay.

VIII. Conclusions

In this paper we investigate the connections between max-weight approaches and dual subgradient methods for convex optimisation. We find that strong connections do indeed exist and we establish a clean, unifying theoretical framework that includes both max-weight and dual subgradient approaches as special cases.

IX. Appendix: Proofs

A. Proof of Theorems 1 and 2

The following two fundamental results are the key to establishing Theorem 1:

Lemma 8. *For any feasible point $\boldsymbol{y} \in C = \text{conv}(D)$ and any vector $\boldsymbol{z} \in \mathbb{R}^n$ there exists a point $\boldsymbol{x} \in D := \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_{|D|}\} \subset \mathbb{R}^n$ such that $\boldsymbol{z}^T(\boldsymbol{x} - \boldsymbol{y}) \le 0$.*

Proof: Since $\boldsymbol{y} \in C = \text{conv}(D)$, $\boldsymbol{y} = \sum_{i=1}^{|D|} a^{(i)}\boldsymbol{x}_i$ with $\sum_{i=1}^{|D|} a^{(i)} = 1$ and $a^{(i)} \in [0, 1]$. Hence, $\boldsymbol{z}^T(\boldsymbol{x} - \boldsymbol{y}) = \sum_{i=1}^{|D|} a^{(i)}\boldsymbol{z}^T(\boldsymbol{x} - \boldsymbol{x}_i)$. Select $\boldsymbol{x} \in \arg\min_{\boldsymbol{w}\in D} \boldsymbol{z}^T\boldsymbol{w}$. Then $\boldsymbol{z}^T\boldsymbol{x} \le \boldsymbol{z}^T\boldsymbol{x}_i$ for all $\boldsymbol{x}_i \in D$ and so $\boldsymbol{z}^T(\boldsymbol{x} - \boldsymbol{y}) \le 0$. ■

Lemma 9 (Non-Convex Descent)**.** *Let $F(\boldsymbol{z})$ be a convex function and suppose points $\boldsymbol{y}$, $\boldsymbol{z} \in C = \text{conv}(D)$ exist such that $F(\boldsymbol{y}) \le F(\boldsymbol{z}) - \epsilon$, $\epsilon > 0$. Suppose $F(\cdot)$ has bounded curvature on C with curvature constant μ_F. Then there exists* *at least one $\boldsymbol{x} \in D$ such that $F((1-\beta)\boldsymbol{z}+\beta\boldsymbol{x}) < F(\boldsymbol{z}) - \gamma\beta\epsilon$ with $0 < \gamma < 1$ provided $\beta \le (1-\gamma)\min\{\epsilon/(\mu_F \bar{x}_D^2), 1\}$.*

Proof: By convexity,

$$F(\boldsymbol{z}) + \partial F(\boldsymbol{z})^T(\boldsymbol{y}-\boldsymbol{z}) \le F(\boldsymbol{y}) \le F(\boldsymbol{z}) - \epsilon.$$

Hence, $\partial F(\boldsymbol{z})^T(\boldsymbol{y}-\boldsymbol{z}) \le -\epsilon$. Now observe that for $\boldsymbol{x} \in D$ we have $(1-\beta)\boldsymbol{z}+\beta\boldsymbol{x} \in C$ and by the bounded curvature of $F(\cdot)$ on C

$$\begin{aligned}
&F((1-\beta)\boldsymbol{z}+\beta\boldsymbol{x}) \\
&\le F(\boldsymbol{z}) + \beta\partial F(\boldsymbol{z})^T(\boldsymbol{x}-\boldsymbol{z}) + \mu_F\beta^2\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 \\
&= F(\boldsymbol{z}) + \beta\partial F(\boldsymbol{z})^T(\boldsymbol{y}-\boldsymbol{z}) + \beta\partial F(\boldsymbol{z})^T(\boldsymbol{x}-\boldsymbol{y}) + \mu_F\beta^2\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 \\
&\le F(\boldsymbol{z}) - \beta\epsilon + \beta\partial F(\boldsymbol{z})^T(\boldsymbol{x}-\boldsymbol{y}) + \mu_F\beta^2\|\boldsymbol{x}-\boldsymbol{z}\|_2^2
\end{aligned}$$

By Lemma 8 we can select $\boldsymbol{x} \in D$ such that $\partial F(\boldsymbol{z})^T(\boldsymbol{x}-\boldsymbol{y}) \le 0$. With this choice of $\boldsymbol{x}$ it follows that

$$\begin{aligned}
F((1-\beta)\boldsymbol{z}+\beta\boldsymbol{x}) &\le F(\boldsymbol{z}) - \beta\epsilon + \mu_F\beta^2\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 \\
&\le F(\boldsymbol{z}) - \beta\epsilon + \mu_F\beta^2\bar{x}_D^2 \qquad (44)
\end{aligned}$$

where (44) follows from Lemma 2, and the result now follows. ∎

Proof of Theorem 1: Since $F_k(\cdot)$ has bounded curvature for any k it is continuous, and as C is closed and bounded we have by the Weierstrass theorem (e.g., see Proposition 2.1.1 in [22]) that $\min_{\boldsymbol{z}\in C} F_k(\boldsymbol{z})$ is finite. We now proceed considering two cases:

Case (i): $F_k(\boldsymbol{z}_k) - F_k(\boldsymbol{y}_k^*) \ge \epsilon$. By Lemma 9 there exists $\boldsymbol{x}_k \in D$ such that $F_k((1-\beta)\boldsymbol{z}_k + \beta\boldsymbol{x}_k) - F_k(\boldsymbol{z}_k) = F_k(\boldsymbol{z}_{k+1}) - F_k(\boldsymbol{z}_k) \le -\gamma\beta\epsilon$. Further, since $F_{k+1}(\boldsymbol{z}_{k+1}) - F_k(\boldsymbol{z}_{k+1}) \le \gamma_1\gamma\beta\epsilon$ and $F_k(\boldsymbol{z}_k) - F_{k+1}(\boldsymbol{z}_k) \le \gamma_1\gamma\beta\epsilon$ it follows

$$F_{k+1}(\boldsymbol{z}_{k+1}) - F_{k+1}(\boldsymbol{z}_k) \le 2\gamma_1\gamma\beta\epsilon - \gamma\beta\epsilon < 0. \qquad (45)$$

That is, $F_k(\cdot)$ and $F_{k+1}(\cdot)$ decrease monotonically when $F_k(\boldsymbol{z}_k) - F_k(\boldsymbol{y}_k^*) \ge \epsilon$.

Case (ii): $F_k(\boldsymbol{z}_k) - F_k(\boldsymbol{y}_k^*) < \epsilon$. It follows that $F_k(\boldsymbol{z}_k) < F_k(\boldsymbol{y}_k^*) + \epsilon$. Since $F_k(\cdot)$ is convex and has bounded curvature, $F_k(\boldsymbol{z}_{k+1}) \le F_k(\boldsymbol{z}_k) + \beta\partial F_k(\boldsymbol{z}_k)^T(\boldsymbol{x}_k - \boldsymbol{z}_k) + \beta^2\mu_F\bar{x}_D^2 \le F_k(\boldsymbol{y}_k^*) + \epsilon + \beta\partial F_k(\boldsymbol{z}_k)^T(\boldsymbol{x}_k - \boldsymbol{z}_k) + \beta^2\mu_F\bar{x}_D^2$. The final term holds uniformly for all $\boldsymbol{x}_k \in D$ and since we select $\boldsymbol{x}_k$ to minimise $F_k(\boldsymbol{z}_{k+1})$ by Lemma 8 we therefore have $F_k(\boldsymbol{z}_{k+1}) \le F_k(\boldsymbol{y}_k^*) + \epsilon + \beta^2\mu_F\bar{x}_D^2$. Using the stated choice of β and the fact that $F_{k+1}(\boldsymbol{z}_{k+1}) - \gamma_1\gamma\beta\epsilon \le F_k(\boldsymbol{z}_{k+1})$ yields

$$F_{k+1}(\boldsymbol{z}_{k+1}) - F_k(\boldsymbol{y}_k^*) \le \gamma_1\gamma\beta\epsilon + \epsilon + \beta(1-\gamma)\epsilon. \qquad (46)$$

Finally, since $F_k(\boldsymbol{y}_k^*) \le F_k(\boldsymbol{y}_{k+1}^*) \le F_{k+1}(\boldsymbol{y}_{k+1}^*) + \gamma_1\gamma\beta\epsilon$ we obtain

$$\begin{aligned}
F_{k+1}(\boldsymbol{z}_{k+1}) - F_{k+1}(\boldsymbol{y}_{k+1}^*) &\le 2\gamma_1\gamma\beta\epsilon + \epsilon + \beta(1-\gamma)\epsilon \\
&\le 2\epsilon.
\end{aligned}$$

We therefore have that $F_{k+1}(\boldsymbol{z}_k)$ is strictly decreasing when $F_k(\boldsymbol{z}_k) - F_k(\boldsymbol{y}_k^*) \ge \epsilon$ and otherwise uniformly upper bounded by 2ϵ. It follows that for all k sufficiently large $F_k(\boldsymbol{z}_{k+1}) - F_k(\boldsymbol{y}_k^*) \le 2\epsilon$ as claimed. ∎

Proof of Theorem 2: Firstly, we make the following observations,

$$\begin{aligned}
&\arg\min_{\boldsymbol{z}\in C} F_k(\boldsymbol{z}_k) + \partial F_k(\boldsymbol{z}_k)^T(\boldsymbol{z}-\boldsymbol{z}_k) \\
&\overset{(a)}{=} \arg\min_{\boldsymbol{z}\in C} \partial F_k(\boldsymbol{z}_k)^T\boldsymbol{z} \overset{(b)}{=} \arg\min_{\boldsymbol{x}\in D} \partial F_k(\boldsymbol{z}_k)^T\boldsymbol{x} \qquad (47)
\end{aligned}$$

where equality (a) follows by dropping terms not involving $\boldsymbol{z}$ and (b) from the observation that we have a linear program (the objective is linear and set C is a polytope, so defined by linear constraints) and so the optimum lies at an extreme point of set C *i.e.*, in set D. We also have that

$$\begin{aligned}
F_k(\boldsymbol{z}_k) + \partial F_k(\boldsymbol{z}_k)^T(\boldsymbol{x}_k - \boldsymbol{z}_k) &\overset{(a)}{\le} F_k(\boldsymbol{z}_k) + \partial F_k(\boldsymbol{z}_k)^T(\boldsymbol{y}_k^* - \boldsymbol{z}_k) \\
&\overset{(b)}{\le} F_k(\boldsymbol{y}_k^*) \le F_k(\boldsymbol{z}_k)
\end{aligned}$$

where $\boldsymbol{y}_k^* \in \arg\min_{\boldsymbol{z}\in C} F_k(\boldsymbol{z})$, inequality (a) follows from the minimality of $\boldsymbol{x}_k$ in C noted above and (b) from the convexity of $F_k(\cdot)$. It follows that $\partial F_k(\boldsymbol{z}_k)^T(\boldsymbol{x}_k - \boldsymbol{z}_k) \le -(F_k(\boldsymbol{z}_k) - F_k(\boldsymbol{y}_k^*)) \le 0$. We have two cases to consider. Case (i): $F_k(\boldsymbol{z}_k) - F_k(\boldsymbol{y}_k^*) \ge \epsilon$. By the bounded curvature of $F_k(\cdot)$,

$$\begin{aligned}
F_k(\boldsymbol{z}_{k+1}) &\le F_k(\boldsymbol{z}_k) + \beta\partial F_k(\boldsymbol{z}_k)^T(\boldsymbol{x}_k - \boldsymbol{z}_k) + \mu_F\beta^2\bar{x}_D \\
&\le F_k(\boldsymbol{z}_k) - \beta\epsilon + \mu_f\beta^2\bar{x}_D \le F_k(\boldsymbol{z}_k) - \gamma\beta\epsilon
\end{aligned}$$

Hence,

$$\begin{aligned}
F_{k+1}(\boldsymbol{z}_{k+1}) &\le F_k(\boldsymbol{z}_{k+1}) + |F_{k+1}(\boldsymbol{z}_{k+1}) - F_k(\boldsymbol{z}_{k+1})| \\
&\le F_k(\boldsymbol{z}_k) - \gamma\beta\epsilon + \gamma_1\gamma\beta\epsilon \le F_k(\boldsymbol{z}_k) - \gamma\beta\epsilon/2,
\end{aligned}$$

and since $F_k(\boldsymbol{z}_k) \le F_{k+1}(\boldsymbol{z}_k) + \gamma_1\gamma\beta\epsilon$ we have that $F_{k+1}(\boldsymbol{z}_{k+1}) - F_k(\boldsymbol{z}_{k+1}) < 0$. Case (ii): $F_k(\boldsymbol{z}_k) - F_k(\boldsymbol{y}_k^*) < \epsilon$. Then

$$\begin{aligned}
F_k(\boldsymbol{z}_{k+1}) &\le F_k(\boldsymbol{z}_k) + \beta\partial F_k(\boldsymbol{z}_k)^T(\boldsymbol{x}_k - \boldsymbol{z}_k) + \mu_F\beta^2\bar{x}_D \\
&\le F_k(\boldsymbol{y}_k^*) + \epsilon + \beta\epsilon,
\end{aligned}$$

and similar to the proof of Theorem 1 we obtain that $F_{k+1}(\boldsymbol{z}_{k+1}) - F_{k+1}(\boldsymbol{y}_{k+1}^*) \le 2\epsilon$. We therefore have that $F_k(\boldsymbol{z}_k)$ is strictly decreasing when $F_k(\boldsymbol{z}_k) - F_k(\boldsymbol{y}_k^*) \ge \epsilon$ and otherwise uniformly upper bounded by 2ϵ. Thus for k sufficiently large $F_k(\boldsymbol{z}_k) - F_k(\boldsymbol{y}_k^*) \le 2\epsilon$. ∎

B. Proof of Lemma 5

Proof of Lemma 5: By Lemma 4 we have it is enough to show that $2\alpha|\sum_{i=1}^k (\boldsymbol{a}^{(j)})^T(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i) + b_k^{(j)} - b^{(j)}|$ is uniformly bounded in order to establish the boundedness of $|\tilde{\lambda}_k^{(j)} - \lambda_k^{(j)}|$ for all $k \ge 1$. Now see that $2\alpha|\sum_{i=1}^k (\boldsymbol{a}^{(j)})^T(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i) + b_k^{(j)} - b^{(j)}| \le 2\alpha|\sum_{i=1}^k (\boldsymbol{a}^{(j)})^T(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i)| + \alpha|\sum_{i=1}^k b_k^{(j)} - b^{(j)}| \le 2\alpha|\sum_{i=1}^k (\boldsymbol{a}^{(j)})^T(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i)| + \alpha\sigma_2$, hence, we only need to show that $|\sum_{i=1}^k (\boldsymbol{a}^{(j)})^T(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i)|$ is bounded.

First of all, since $\boldsymbol{z}_{i+1} = (1-\beta)\boldsymbol{z}_i + \beta\boldsymbol{x}_i$ we have $\boldsymbol{z}_{i+1} - \boldsymbol{x}_i = (1-\beta)(\boldsymbol{z}_i - \boldsymbol{x}_i)$. Then, $\sum_{i=1}^k(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i) = (1-\beta)\sum_{i=1}^k(\boldsymbol{z}_i - \boldsymbol{x}_i)$. Using again the fact that $\boldsymbol{z}_i - \boldsymbol{x}_{i-1} = (1-\beta)(\boldsymbol{z}_{i-1} - \boldsymbol{x}_{i-1})$ it follows that $\sum_{i=1}^k(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i) = (1-\beta)(\boldsymbol{z}_1 - \boldsymbol{x}_k) + \sum_{i=1}^{k-1}(1-\beta)^2(\boldsymbol{z}_i - \boldsymbol{x}_i)$. Applying the preceding argument recursively we obtain that $\sum_{i=1}^k(\boldsymbol{z}_{i+1} -$

$\boldsymbol{x}_i) = (1-\beta)(\boldsymbol{z}_1 - \boldsymbol{x}_k) + (1-\beta)^2(\boldsymbol{z}_1 - \boldsymbol{x}_{k-1}) + \cdots + (1-\beta)^k(\boldsymbol{z}_1 - \boldsymbol{x}_1) + (1-\beta)^{k+1}(\boldsymbol{z}_1 - \boldsymbol{x}_1)$, and so

$$\sum_{i=1}^{k}(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i) = \sum_{i=1}^{k}(1-\beta)^{k+1-i}(\boldsymbol{z}_1 - \boldsymbol{x}_i). \tag{48}$$

Using (48) it follows that

$$\begin{aligned} 2\alpha & \left|\sum_{i=1}^{k}(\boldsymbol{a}^{(j)})^T(\boldsymbol{z}_{i+1} - \boldsymbol{x}_i)\right| \\ & \leq 2\alpha\left|\sum_{i=1}^{k}(1-\beta)^{k+1-i}(\boldsymbol{a}^{(j)})^T(\boldsymbol{z}_1 - \boldsymbol{x}_i)\right| \\ & \leq 2\alpha\sigma_1\sum_{i=1}^{k}(1-\beta)^{k+1-i} \end{aligned} \tag{49}$$

where $\sigma_1 = \max_{j\in\{1,\ldots,m\}}\max_{\boldsymbol{x}\in D,\boldsymbol{z}\in C}|(\boldsymbol{a}^{(j)})^T(\boldsymbol{z}-\boldsymbol{x})|$. Next, see that $\sum_{i=1}^{k}(1-\beta)^{k+1-i} = (1-\beta)^{k+1}\sum_{i=1}^{k}(1-\beta)^{-i}$ and that

$$\sum_{i=1}^{k}\frac{1}{(1-\beta)^i} = \frac{1-(1-\beta)^{k+1}}{\beta(1-\beta)^k},$$

therefore,

$$(1-\beta)^{k+1}\sum_{i=1}^{k}(1-\beta)^{-i} \leq \frac{1-(1-\beta)^{k+2}}{\beta} \leq \frac{1}{\beta}.$$

Finally, using the latter bound in (49) the stated result now follows. ■

References

[1] A. Stolyar, "Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm," *Queueing Systems*, vol. 50, no. 4, pp. 401–457, 2005.

[2] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *Automatic Control, IEEE Transactions on*, vol. 37, no. 12, pp. 1936–1948, Dec 1992.

[3] M. Neely, E. Modiano, and C. Rohrs, "Power allocation and routing in multibeam satellites with time-varying channels," *Networking, IEEE/ACM Transactions on*, vol. 11, no. 1, pp. 138–152, Feb 2003.

[4] 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, pp. 1333–1344, Dec 2007.

[5] A. L. Stolyar, "Greedy primal-dual algorithm for dynamic resource allocation in complex networks," *Queueing Systems*, vol. 54, no. 3, pp. 203–220, 2006.

[6] M. Neely, E. Modiano, and C. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 23, no. 1, pp. 89–103, Jan 2005.

[7] M. Neely, "Energy optimal control for time-varying wireless networks," *Information Theory, IEEE Transactions on*, vol. 52, no. 7, pp. 2915–2934, July 2006.

[8] M. Neely, E. Modiano, and C. ping Li, "Fairness and optimal stochastic control for heterogeneous networks," *Networking, IEEE/ACM Transactions on*, vol. 16, no. 2, pp. 396–409, April 2008.

[9] M. Neely, *Stochastic network optimization with application to communication and queueing systems*. Morgan & Claypool Publishers, 2010.

[10] L. Georgiadis, M. Neely, and L. Tassiulas, *Resource Allocation and Cross-Layer Control in Wireless Networks*. Morgan & Claypool Publishers, 2006.

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

[12] X. Lin, N. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Selected Areas in Communications*, vol. 24, no. 8, pp. 1452–1463, 2006.

[13] L. Huang and M. Neely, "Delay reduction via lagrange multipliers in stochastic network optimization," *IEEE Trans Automatic Control*, vol. 56, no. 4, pp. 842–857, 2011.

[14] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[15] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., 1989.

[16] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.

[17] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *ICML*, 2003, pp. 928–936.

[18] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '05. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2005, pp. 385–394. [Online]. Available: http://dl.acm.org/citation.cfm?id=1070432.1070486

[19] A. Roberts and D. Varberg, "Another proof that convex functions are locally lipschitz," *The American Mathematical Monthly*, vol. 81, no. 9, pp. 1014–1016, Nov 1974.

[20] S. P. Meyn, *Control techniques for complex networks*. Cambridge University Press, 2008.

[21] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of Optimization Theory and Applications*, vol. 142, no. 1, pp. 205–228, 2009.

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