Dual Subgradient Methods Using Approximate Multipliers

Víctor Valls, Douglas J. Leith
Trinity College Dublin

***Abstract*—We consider the subgradient method for the dual problem in convex optimisation with approximate multipliers, *i.e.*, the subgradient used in the update of the dual variables is obtained using an approximation of the true Lagrange multipliers. This problem is interesting for optimisation problems where the exact Lagrange multipliers might not be readily accessible. For example, in distributed optimisation the exact Lagrange multipliers might not be available at the nodes due to communication delays or losses. We show that we can construct approximate primal solutions that can get arbitrarily close to the set of optima as step size α is reduced. Applications of the analysis include unsynchronised subgradient updates in the dual variable update and unsynchronised max-weight scheduling.**

***Index Terms*—subgradient methods, Lagrange multipliers, max-weight scheduling, unsynchronised updates.**

I. Introduction

In this paper we study the convergence of the subgradient method with constant step size in constrained convex optimisation with *approximate* subgradients. More specifically, the subgradient used in the update of the dual variables is computed using an approximation of the Lagrange multipliers instead of the *true* Lagrange multipliers.

One of the motivations of considering approximate multipliers in the optimisation is that in some problems the exact Lagrange multipliers might not be available, but a *noisy*, *delayed* or *perturbed* version of the multipliers is available instead. For instance, in distributed optimisation a node might not have access to the exact Lagrange multiplier in the system due to transmission delays or losses. Another example is in network problems where discrete valued queue occupancies can be identified with approximate scaled Lagrange multipliers [1].

We show that the subgradient method with approximate multipliers converges as long as the distance between the approximate and Lagrange multipliers can be controlled with step size $\alpha > 0$. In particular, we show that the running average of the associated primal points can get arbitrarily close to the set of primal optima as step size α is reduced. Further, this running average is asymptotically attracted to a feasible point for every $\alpha > 0$.

These results establish conditions under which inaccurate knowledge of the Lagrange multipliers does not affect the convergence of the subgradient method. Important applications of the analysis include unsynchronised subgradient updates of the dual variable and unsynchronised max-weight scheduling.

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

A. Related Work

Subgradient methods in constrained convex optimisation have been extensively studied under various steps size rules. For example, see the early works by Shor [2] and Polyak [3] or the more modern works by Bertsekas *et. al.* in [4] and [5].

Subgradient methods in constrained convex optimisation with noisy updates have been studied by Nedić and Bertsekas in [6]. In particular, the authors study the case where the noise is deterministic and bounded, and show optimality using several step size rules. The work in [7] considers unsynchronised subgradient updates where a subset of subgradients are updated at each iteration. The convergence of the algorithm is proved using a diminishing step size rule. The work in [8] considers asynchronous subgradient updates with constant step size, and asymptotic optimality is proved using averaging schemes. Approximate solutions to convex problems under an averaging scheme have been studied by Nedić in [9] and [10]. The work in [9] assumes that the dual function can be computed efficiently and the work in [10] considers a sequence of primal-dual subgradient updates. For a good reference on primal averaging schemes see the related work section and references in [9].

The max-weight scheduling introduced by Tassiulas in [11] has been extended in a wide range of papers to consider the minimise a convex utility function. For instance, energy minimisation [12] or fairneess [13] among others. In [1] we showed that Lagrange multipliers and queue occupancies in network are related, and that max-weight resource allocation problems can be formulated as convex optimisation problems.

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 $x^{(j)}$ to denote the j'th element of a vector $\boldsymbol{x}$. The j'th element of operator $[\boldsymbol{x}]^+$ equals $x^{(j)}$ (the j'th element of $\boldsymbol{x}$) when $x^{(j)} > 0$ and otherwise equals 0 when $x^{(j)} < 0$. 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 $y^{(j)} \geq x^{(j)}$, $y^{(j)} > x^{(j)}$ respectively for all $j = 1, \dots, m$.

II. Preliminaries

A. Problem Setup

Consider the following convex optimisation problem P:

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

where $f, g^{(j)} : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \dots, m$ are convex functions, $\boldsymbol{g}(\boldsymbol{z}) = [g^{(1)}(\boldsymbol{z}), \dots, g^{(m)}(\boldsymbol{z})]^T$ and C is a bounded convex set from $\mathbb{R}^n$. Let $C_0 := \{\boldsymbol{z} \in C \mid \boldsymbol{g}(\boldsymbol{z}) \preceq \mathbf{0}\}$ and we will assume C_0 is non-empty, *i.e.*, problem P is feasible. Further, we will denote by $C^\star := \arg\min_{\boldsymbol{z}\in C_0} f(\boldsymbol{z}) \subseteq C_0$ the set of optima and $f^\star := f(\boldsymbol{z}^\star)$, $\boldsymbol{z}^\star \in C^\star$.

B. Lagrange Penalty & Dual Function

As in classical constrained convex optimisation we define the Lagrange penalty function $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated with problem P as

$$L(\boldsymbol{z}, \boldsymbol{\lambda}) := f(\boldsymbol{z}) + \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{z}) \tag{1}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m_+$. Note that $L(\boldsymbol{z}, \boldsymbol{\lambda})$ is convex in $\boldsymbol{z}$ for a fixed $\boldsymbol{\lambda}$ and linear in $\boldsymbol{\lambda}$ for a fixed $\boldsymbol{z}$. We also define the Lagrange dual function as follows

$$q(\boldsymbol{\lambda}) := L(\boldsymbol{z}^*(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = \min_{\boldsymbol{z}\in C} L(\boldsymbol{z}, \boldsymbol{\lambda}) \qquad \boldsymbol{\lambda} \succeq \mathbf{0}$$

where $\boldsymbol{z}^*(\boldsymbol{\lambda}) \in \arg\min_{\boldsymbol{z}\in C} L(\boldsymbol{z}, \boldsymbol{\lambda})$. Importantly, since $q(\boldsymbol{\lambda})$ is the infimum of a collection of linear functions it is concave and $q(\boldsymbol{\lambda}) \le L(\boldsymbol{z}, \boldsymbol{\lambda})$ for all $\boldsymbol{\lambda} \succeq \mathbf{0}$. The dual function is of particular interest when we have strong duality, *i.e.*, the solution of the dual problem P^D

$$\underset{\boldsymbol{\lambda}\succeq\mathbf{0}}{\text{maximise}} \quad q(\boldsymbol{\lambda}) \tag{2}$$

and the primal problem P coincide. Namely, when we have that

$$f^\star = \min_{\boldsymbol{z}\in C} \max_{\boldsymbol{\lambda}\succeq\mathbf{0}} L(\boldsymbol{z}, \boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}\succeq\mathbf{0}} \min_{\boldsymbol{z}\in C} L(\boldsymbol{z}, \boldsymbol{\lambda}) = q(\boldsymbol{\lambda}^\star)$$

where $\boldsymbol{\lambda}^\star := \arg\max_{\boldsymbol{\lambda}\succeq\mathbf{0}} q(\boldsymbol{\lambda})$. We are guaranteed to have strong duality when the following assumption holds.

Assumption 1 (Slater condition)**.** *The set $C_0 := \{\boldsymbol{z} \in C \mid \boldsymbol{g}(\boldsymbol{z}) \preceq \mathbf{0}\}$ has non-empty relative interior,* i.e.*, there exists a point $\bar{\boldsymbol{z}} \in C$ such that $\boldsymbol{g}(\bar{\boldsymbol{z}}) \prec \mathbf{0}$.*

A direct consequence of the Assumption 1 is that the set of dual optima is bounded. The following lemma corresponds to Lemma 1 in [9].

Lemma 1 (Bounded Dual Set)**.** *Let $Q_\delta := \{\boldsymbol{\lambda} \succeq \mathbf{0} : q(\boldsymbol{\lambda}) \ge q(\boldsymbol{\lambda}^\star) - \delta\}$ with $\delta \ge 0$ and let the Slater condition hold,* i.e.*, there exists a vector $\bar{\boldsymbol{z}} \in C$ such that $\boldsymbol{g}(\bar{\boldsymbol{z}}) \prec \mathbf{0}$. Then, for every $\boldsymbol{\lambda} \in Q_\delta$ we have that*

$$\|\boldsymbol{\lambda}\|_2 \le \mathcal{Q} := \frac{1}{\upsilon}(f(\bar{\boldsymbol{z}}) - f^\star + \delta) \tag{3}$$

where $\upsilon := \min_{j\in\{1,\dots,m\}} -g^{(j)}(\bar{\boldsymbol{z}})$.

C. Classical Subgradient Method

Problem P^D is an unconstrained optimisation problem that can be solved, for example, using the subgradient method. The subgradient method for the dual problem consists of updates

$$\boldsymbol{z}_k^*(\boldsymbol{\lambda}_k) \in \arg\min_{\boldsymbol{z}\in C} L(\boldsymbol{z}, \boldsymbol{\lambda}_k) \tag{4}$$

$$\boldsymbol{\lambda}_{k+1} = [\boldsymbol{\lambda}_k + \alpha(\boldsymbol{g}(\boldsymbol{z}_k^*(\boldsymbol{\lambda}_k)))]^+ \tag{5}$$

where $\alpha > 0$ is a step size. Subgradient methods can make use of more complex step sizes (see [5] for example), however, in this paper we will only use a constant step size for its simplicity and practical importance.

The convergence of the subgradient method for the dual problem [4] is based on decreasing the euclidean distance between $\boldsymbol{\lambda}_k$ and $\boldsymbol{\lambda}^\star$. In particular, the distance between $\boldsymbol{\lambda}_k$ and $\boldsymbol{\lambda}^\star$ decreases when the difference between $q(\boldsymbol{\lambda}_k)$ and $q(\boldsymbol{\lambda}^\star)$ is sufficiently large; otherwise $\boldsymbol{\lambda}_k$ remains in a ball around $\boldsymbol{\lambda}^\star$. The size of the ball $\boldsymbol{\lambda}_k$ converges depends on step size $\alpha > 0$, having that $\boldsymbol{\lambda}_k \to \Lambda^\star := \arg\max_{\boldsymbol{\lambda}\succeq\mathbf{0}} q(\boldsymbol{\lambda})$ as $\alpha \to 0$. Further, since by strong duality arguments $\boldsymbol{z}^*(\boldsymbol{\lambda}_k) \to \boldsymbol{z}^\star$ as $\alpha \to 0$ we can intuitively expect $\boldsymbol{z}^*(\boldsymbol{\lambda}_k)$ to converge *close* to $\boldsymbol{z}^\star \in C^\star$ for α sufficiently small.

III. Subgradient Method With Approximate Multipliers

The classical subgradient updates (4) and (5) can be extended to make use of approximate Lagrange multipliers $\boldsymbol{\mu} \in \mathbb{R}^m_+$. Using an approximation of the Lagrange multiplies in the subgradient method is interesting, for example, to capture the fact that in some optimisation problems the exact Lagrange multipliers might not be known or have errors.

Suppose we have the following updates

$$\boldsymbol{z}_k^*(\boldsymbol{\mu}_k) \in \arg\min_{\boldsymbol{z}\in C} L(\boldsymbol{z}, \boldsymbol{\mu}_k) \tag{6}$$

$$\boldsymbol{\lambda}_{k+1} = [\boldsymbol{\lambda}_k + \alpha \boldsymbol{g}(\boldsymbol{z}_k^*(\boldsymbol{\mu}_k))]^+ \tag{7}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^m_+$. We can think of $\boldsymbol{\mu}_k$ as being an approximate multiplier substituted into the classical subgradient updates (4) and (5). We are first interested in knowing when the multiplier $\boldsymbol{\lambda}_k$ generated by (7) will converge to a ball around $\boldsymbol{\lambda}^\star$. This is answered by the following lemma.

Lemma 2. *Consider the setup of problem P and updates (6) and (7) and suppose that $\|\boldsymbol{\lambda}_k - \boldsymbol{\mu}_k\|_\infty \le \alpha\sigma_0$ for all k with $\alpha > 0$, $\sigma_0 \ge 0$. Suppose the Slater condition is satisfied (Assumption 1) and that $\boldsymbol{\lambda}_1 \in \mathbb{R}^m_+$. Then $\boldsymbol{\lambda}_k$ converges to a ball around $\boldsymbol{\lambda}^\star$. Further, the size of the ball depends on parameter $\alpha > 0$, having that $\boldsymbol{\lambda}_k \to \Lambda^\star := \arg\max_{\boldsymbol{\lambda}\succeq\mathbf{0}} q(\boldsymbol{\lambda})$ as $\alpha \to 0$.*

Lemma 2 tells us that all we require for $\boldsymbol{\lambda}_k$ to converge to a ball around $\boldsymbol{\lambda}^\star$ is that the distance between $\boldsymbol{\mu}_k$ and $\boldsymbol{\lambda}_k$ is uniformly bounded for all k, *i.e.*, $\|\boldsymbol{\lambda}_k - \boldsymbol{\mu}_k\|_\infty \le \alpha\sigma_0$ with $\alpha, \sigma_0 > 0$. As we will see later, this condition can be readily evaluated and is commonly satisfied in practical problems of interest. Further, observe that when $\sigma_0 = 0$ then $\boldsymbol{\mu}_k = \boldsymbol{\lambda}_k$ and so we recover the classical subgradient method.

In addition to convergence of the multipliers, we also have that the running average

$$\boldsymbol{z}_k^\diamond := \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{z}_i^*(\boldsymbol{\mu}_i) \tag{8}$$

converges to a feasible point as $k \to \infty$ for every $\alpha > 0$. This is formally presented in the following lemma.

Lemma 3. *(Feasibility) Consider the setup of problem P and updates (6) and (7) where $\|\boldsymbol{\lambda}_k - \boldsymbol{\mu}_k\|_\infty \leq \alpha\sigma_0$ for all k with $\alpha > 0$, $\sigma_0 \geq 0$. Suppose the Slater condition is satisfied (Assumption 1) and that $\boldsymbol{\lambda}_1 \in \mathbb{R}^m_+$. Then, for all $k = 1, 2, \dots$*

$$g(\boldsymbol{z}_k^\diamond) := g\Big(\frac{1}{k}\sum_{i=1}^{k} \boldsymbol{z}_i^*(\boldsymbol{\mu}_i)\Big) \leq \frac{\bar{\lambda}}{\alpha k} \tag{9}$$

where $\|\boldsymbol{\lambda}_k\|_2 \leq \bar{\lambda} := 2\mathcal{Q} + \max\{\|\boldsymbol{\lambda}_1\|_2, \mathcal{Q} + \alpha m \bar{g}\}$, $\bar{g} := \max_{\boldsymbol{z} \in C} \|\boldsymbol{g}(\boldsymbol{z})\|_\infty$ and $\mathcal{Q}$ is given in Lemma 1 with $\delta := \alpha m^2(\bar{g}^2/2 + \sigma_0\bar{g})$. As $k \to \infty$ it follows that $g(\boldsymbol{z}_k^\diamond) \leq 0$ and so $\boldsymbol{z}_k^\diamond$ is feasible.

Also, by strong duality arguments we can expect $\boldsymbol{z}_k^\diamond$ to converge to a feasible point that is *close* to $\boldsymbol{z}^\star$ for α sufficiently small. This intuition is captured by our main result, which establishes the convergence of $f(\boldsymbol{z}_k^\diamond)$ to $f^\star$ depending on step size α and iterate k.

Theorem 1. *Consider the setup of problem P and updates (6) and (7) where $\|\boldsymbol{\lambda}_k - \boldsymbol{\mu}_k\|_\infty \leq \alpha\sigma_0$ for all $j = 1, \dots, m$ with $\alpha > 0$, $\sigma_0 \geq 0$. Suppose the Slater condition is satisfied (Assumption 1). Then,*

$$-\frac{2m\bar{\lambda}^2}{\alpha k} - \frac{\alpha}{2}m(\bar{g}^2/2 + \sigma_0\bar{g}) \leq f(\boldsymbol{z}_k^\diamond) - f^\star \leq \alpha m(\bar{g}^2 + \sigma_0\bar{g}) + \frac{3m\bar{\lambda}^2}{2\alpha k} \tag{10}$$

Proof: See appendix. ■

Observe that the bounds in Theorem 1 are not asymptotic but rather can be applied at finite times k. As $k \to \infty$ the bounds simplify to $-\frac{\alpha}{2}m^2(\bar{g}^2 + 2\sigma_0\bar{g}) \leq f(\boldsymbol{z}_k^\diamond) - f^\star \leq \alpha m^2(\bar{g}^2 + \sigma_0\bar{g})$ and can be made small by selecting step size α sufficiently small. Note from Theorem 1 that if the Lagrange multipliers are accessible in the optimisation ($\boldsymbol{\mu}_k = \boldsymbol{\lambda}_k$ for all k) then $\sigma_0 = 0$ and so we have a tighter bound.

IV. Unsynchronised Updates

Theorem 1 has many useful applications. In addition to allowing use of "noisy" multipliers *e.g.*, $\mu_k^{(j)} = [\lambda_k^{(j)} + \alpha y_k]^+$ where y_k is a bounded random variable, in this section we show that convergence of the subgradient method with unsynchronised updates can be obtained as a corollary, avoiding the need for complex, specialised analysis machinery.

To see this consider the following separable convex problem P^s:

$$\begin{aligned} \underset{\substack{z^{(j)} \in C_j,\\ j=1,\dots,n}}{\text{minimise}} \quad & f(\boldsymbol{z}) := \sum_{j=1}^{n} f^{(j)}(z^{(j)}) \\ \text{subject to} \quad & g_r(\boldsymbol{z}) := \sum_{j=1}^{n} g_r^{(j)}(z^{(j)}) \leq 0 \quad r = 1, \dots, m \end{aligned}$$

where $f^{(j)}, g_r^{(j)} : \mathbb{R} \to \mathbb{R}$, $j = 1, \dots, n$, $r = 1, \dots, m$ are convex functions and $C := C_1 \times \cdots \times C_n$. Here, the objective can be decomposed into functions $f^{(j)}$ which depend on element $z^{(j)}$ of vector $\boldsymbol{z}$. Each element $z^{(j)}$ is constrained to lie in convex set C_j. The constraints $g_r(\boldsymbol{z})$ can also be decomposed but are not separable since the *sum* on the LHS must satisfy the non-positivity constraint (which imposes a joint constraint on the $g_r^{(j)}$).

The Lagrangian of problem P^s is given by

$$\begin{aligned} L(\boldsymbol{z}, \boldsymbol{\lambda}) &= \sum_{j=1}^{n} f^{(j)}(z^{(j)}) + \sum_{r=1}^{m} \lambda^{(r)} \sum_{j=1}^{n} g_r^{(j)}(z^{(j)}) \\ &= \sum_{j=1}^{n} L^{(j)}(z^{(j)}, \boldsymbol{\lambda}) \end{aligned}$$

where $L^{(j)}(z^{(j)}, \boldsymbol{\lambda}) = f^{(j)}(z^{(j)}) + \sum_{r=1}^{m} \lambda^{(r)} g_r^{(j)}(z^{(j)})$ and $\boldsymbol{\lambda} := [\lambda^{(1)}, \dots, \lambda^{(m)}]^T$. The dual function can be decomposed as the sum of n concave functions

$$q(\boldsymbol{\lambda}) := \min_{\substack{z^{(j)} \in C_j,\\ j=1,\dots,n}} \sum_{j=1}^{n} L^{(j)}(z^{(j)}, \boldsymbol{\lambda}) = \sum_{j=1}^{n} q^{(j)}(\boldsymbol{\lambda}). \tag{11}$$

where $q^{(j)}(\boldsymbol{\lambda}) := \min_{z^{(j)} \in C_j} L^{(j)}(z^{(j)}, \boldsymbol{\lambda})$.

While in (11) each function $q^{(j)}$ uses the same multiplier $\boldsymbol{\lambda}$, this is not essential. Instead, we can allow each function $q^{(j)}$ to use a different approximate multiplier $\boldsymbol{\mu}^{(j)}$ and select

$$z_k^{(j)}(\boldsymbol{\mu}_k^{(j)}) \in \arg\min_{w \in C_j} L^{(j)}(w, \boldsymbol{\mu}_k^{(j)}) \qquad j = 1, \dots, n \tag{12}$$

$$\lambda_{k+1}^{(r)} = \Big[\lambda_k^{(r)} + \alpha \sum_{j=1}^{n} g_r^{(j)}(z^{(j)}(\boldsymbol{\mu}_k^{(j)}))\Big]^+ \quad r = 1, \dots, m. \tag{13}$$

Provided $\|\boldsymbol{\lambda}_k - \boldsymbol{\mu}_k^{(j)}\|_\infty \leq \alpha\sigma_0$ for all k, then by Theorem 1 we know that the updates (12)-(13) converge to a ball around the optimum.

In particular, consider now the situation where at each step k only a subset of the elements $z^{(j)}$, $j = 1, \cdots, n$ are updated *i.e.* updates are unsynchronised. The elements which are updated use multiplier $\boldsymbol{\mu}_k^{(j)} = \boldsymbol{\lambda}_k^{(j)}$. Those elements $z^{(j)}$ which are not updated at step k can be formally thought of as being updated using the old multiplier value $\boldsymbol{\mu}_k^{(j)} = \boldsymbol{\lambda}_{k-\tau_k^{(j)}}^{(j)}$ where $k - \tau_k^{(j)}$ is the time step where $z^{(j)}$ was last updated. Provided $\tau_k^{(j)}$ is uniformly upper bounded by σ_1 then $\|\boldsymbol{\lambda}_k - \boldsymbol{\mu}_k^{(j)}\|_\infty$ is unformly bounded by $\alpha\sigma_1\bar{g}$ and so by Theorem 1 updates (12)-(13) converge to a ball around the optimum. That is, we have the following corollary.

Corollary 1. *(Unsynchronised Subgradient Updates) Consider the setup of problem P^s and let $\boldsymbol{\mu}_k^{(j)} = \boldsymbol{\lambda}_{k-\tau_k^{(j)}}^{(j)}$, $j = 1, \dots, n$ with $\boldsymbol{\tau}_k \in \mathbb{N}^n$. Then, the bound (10) holds if there exists a constant $\sigma_1 \geq 0$ such that $\|\boldsymbol{\tau}_k\|_\infty \leq \sigma_1$ for all $k \geq 1$.*

Note that the delay $\tau_k^{(j)}$ may vary from time step to time step and from element to element in an arbitrary manner (including randomly) and this analysis still applies provided that $\tau_k^{(j)}$ is uniformly upper bounded. Extension to random delays which are not uniformly bounded can be made on a sample path basis. That is, we now have a probability of convergence with the probability being the fraction of sample paths for which $\tau_k^{(j)}$ satisfies a specified uniform upper bound. In this way the convergence of many consensus-like approaches for distributed optimisation may be encompassed by Theorem 1.

Further, observe that alternatively each element $z^{(j)}$, $j = 1, \cdots, n$ might be updated using a delayed value of multiplier $\boldsymbol{\lambda}_k^{(j)}$ (these might, for example, be due to propagation delay across a network) and the same analysis applies.

To illustrate these results we present a brief example with unsynchronised updates and approximate multipliers and compare it to the classical subgradient method ($\sigma_0 = 0$ and synchronised subgradient updates).

Example 1. Consider the following optimisation problem

$$\begin{aligned} \underset{\boldsymbol{z}\in C}{\text{minimise}} \quad & \|\boldsymbol{z} - \boldsymbol{w}\|_2^2 \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{z} \preceq \boldsymbol{b} \end{aligned}$$

where $\boldsymbol{z} \in \mathbb{R}^3$, $C := \{\boldsymbol{z} \in \mathbb{R}^3 \mid \boldsymbol{0} \preceq \boldsymbol{z} \preceq \boldsymbol{1}\}$,

$$\boldsymbol{w} := \begin{bmatrix} 1 & \frac{3}{2} & 1 \end{bmatrix}^T, \ \boldsymbol{A} := \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \ \text{and} \ \boldsymbol{b} := \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \end{bmatrix}^T.$$

The Lagrangian is

$$\begin{aligned} L(\boldsymbol{z}, \boldsymbol{\lambda}) &= \|\boldsymbol{z} - \boldsymbol{w}\|_2^2 + \boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{z} - \boldsymbol{b}) \\ &= -\boldsymbol{\lambda}^T \boldsymbol{b} + \sum_{j=1}^{3} (z^{(j)} - w^{(j)})^2 + (\boldsymbol{\lambda}^T \boldsymbol{A}^{(j)}) z^{(j)}, \end{aligned}$$

where $\boldsymbol{\lambda} := [\lambda^{(1)}, \lambda^{(2)}]^T$ and the dual variable updates are given by

$$\lambda_{k+1}^{(j)} = [\lambda_k^{(j)} + \alpha(\boldsymbol{A}\boldsymbol{z}_k^*(\boldsymbol{\mu}_k) - \boldsymbol{b})]^+ \qquad j = 1, 2.$$

We study two cases. Case (i): $\mu_k^{(j)} = \lambda_k^{(j)}$ for all k, $j = 1, 2$ and the subgradients in the dual variable update are computed synchronously. Namely, at each iterate k we have

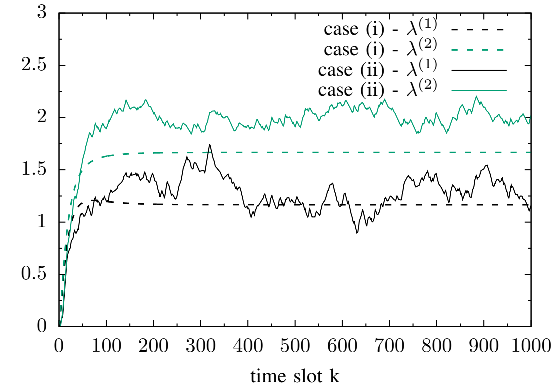
$$z_k^{*(j)} \in \arg \min_{\boldsymbol{s}\in[0,1]} (s - w^{(j)})^2 + (\boldsymbol{\lambda}_k^T \boldsymbol{A}^{(j)}) s$$

for all $j = 1, 2, 3$. Case (ii): approximate multipliers are given by $\mu_k^{(j)} = [\lambda_k^{(j)} + \alpha y_k]^+$ where y_k is a realisation of a random variable uniformly distributed between -1 and 1, and the primal variables $z^{(j)}$ are updated in an unsynchronised manner, *i.e.*,

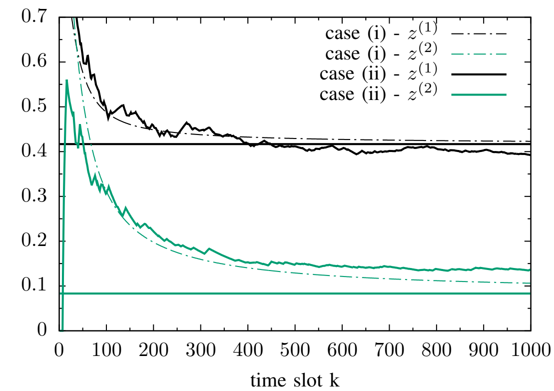
$$z_k^{*(j)} \in \arg \min_{\boldsymbol{s}\in[0,1]} (s - w^{(j)})^2 + (\boldsymbol{\mu}_k^T \boldsymbol{A}^{(j)}) s$$

where $j \in \{1, 2, 3\}$ is selected uniformly at random at each time step. Note that case (i) corresponds to the classical subgradient method for the dual problem and that case (ii) extends it to make use of approximate Lagrange multipliers and unsynchronised updates.

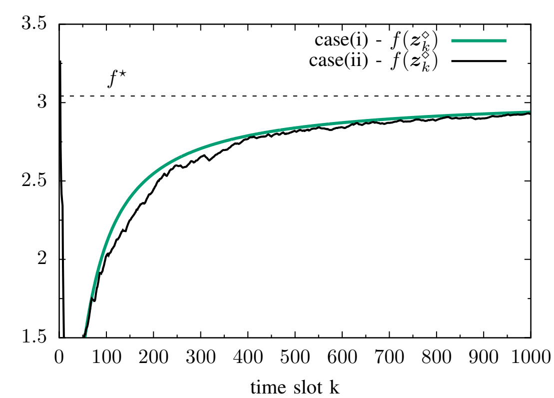
We numerically solve this example with initial condition $\boldsymbol{\lambda}_1 = \boldsymbol{0}$ and step size $\alpha = 1/25$ for both cases. Figures 1a, 1b and 1c show respectively the convergence of $\boldsymbol{\lambda}_k$, $\boldsymbol{z}_k^*(\boldsymbol{\mu}_k)$, $f(\boldsymbol{z}_k^\diamond)$ to $\boldsymbol{\lambda}^\star$, $\boldsymbol{z}^\star$, $f^\star$. Observe from Figure 1a that even though $\boldsymbol{\lambda}_k$ converges to a bigger ball around $\boldsymbol{\lambda}^\star$ in case (ii) than in case (i), the value of $f(\boldsymbol{z}_k^\diamond)$ obtained in both cases is very close for k large enough.



(a) Convergence of the Lagrange multipliers. Optimal dual variables are $\lambda^{\star(1)} = 1.17$ and $\lambda^{\star(2)} = 1.67$.



(b) Convergence of primal variables $z^{(j)}$ to $z^{*(j)}$ for $j = 1, 2$. Straight lines corresponds to optimal values.



(c) Convergence of the objective function in both cases to $f^\star = 3.04$.

Fig. 1: Illustrating the convergence of Example 1 with parameters $\boldsymbol{\lambda}_1 = \boldsymbol{0}$ and $\alpha = 1/25$.

V. Unsynchronised Max-Weight

It was recently shown in [1] and [14] that max-weight scheduling uses the queues within a network as approximate scaled multipliers. This observation can be viewed as essentially a special case of Theorem 1 which makes use of the fact that two queue-like updates $Q_{k+1} = [Q_k + \alpha\delta_k]^+$ and

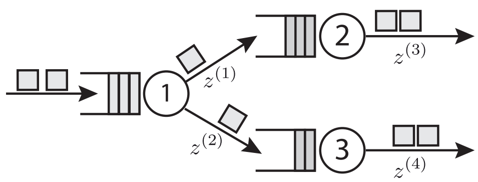


Fig. 2: Network of the example in Section V.

$\tilde{Q}_{k+1} = [\tilde{Q}_k + \alpha\tilde{\delta}_k]^+$ satisfy $|Q_k - \tilde{Q}_k| \leq \alpha\sigma_0$ whenever $\max_k |\sum_{j=1}^k \delta_j - \tilde{\delta}_j| \leq \sigma_0/2$ (a property which also follows from the continuity of queues in the Skorohod metric). Using Theorem 1, the extension to unsynchronised updates is immediate. We illustrate this via a simple network flow example.

A. Network Flow Example

Consider the network illustrated in Figure 2. Time is slotted at each node but nodes' time slots are not synchronised. Exogenous packets arrive into the system (enter the queue of node 1) with mean rate $\frac{1}{2}$ packet per slot. Node 1 can transmit a single packet to either node 2 or node 3; the packets transmitted by nodes 2 and 3 leave the system. All nodes in the network transmit a single packet in a slot.

Similarly to max-weight scheduling, we aim to design a packet scheduling policy at each node that keeps the system stable (bounded queues), does not require us to know the traffic characteristics in advance and makes decisions based only on queues occupancies. Furthermore, nodes make scheduling decisions in order to minimise a local convex utility function of the average link throughput.

B. Fluid-like Optimisation

We can formulate the resource allocation problem as a convex optimisation

$$\min_{\boldsymbol{z}\in C} \sum_{j=1}^{4} f^{(j)}(z^{(j)}) \tag{14}$$
$$\text{s.t.} \quad \frac{1}{2} \leq z^{(1)} + z^{(2)},\ z^{(1)} \leq z^{(3)},\ z^{(2)} \leq z^{(4)}$$

where $C := \{\boldsymbol{z} \in \mathbb{R}^4_+ \mid \|\boldsymbol{z}\|_\infty \leq 1, z^{(1)} + z^{(2)} \leq 1\}$, i.e., it is the convex hull of actions set $\{0, 1\}$ (where 1 represents transmitting a packet over a link whereas 0 to doing nothing) with the restriction that node 1 can only transmit in only one link in a time slot. Constraint $\frac{1}{2} \leq z^{(1)}+z^{(2)}$ enforces the mean service rate of node 1 to be bigger or equal than the mean arrival rate in the system. Similarly, constraints $z^{(1)} \leq z^{(3)}$ and $z^{(2)} \leq z^{(4)}$ enforce the service rate of nodes 2 and 3 to be bigger or equal than the mean service rate of node 1 over each link.

The Lagrangian is given by

$$\begin{aligned} L(\boldsymbol{z}, \boldsymbol{\lambda}) = &\sum_{j=1}^{4} f^{(j)}(z^{(j)}) + \lambda^{(1)}(\frac{1}{2} - z^{(1)} - z^{(2)}) \\ &+ \lambda^{(2)}(z^{(1)} - z^{(3)}) + \lambda^{(3)}(z^{(2)} - z^{(4)}). \end{aligned} \tag{15}$$

Now consider the following subgradient updates with approximate multipliers

$$\begin{aligned} [z_k^{*(1)}(\boldsymbol{\mu}_k), z_k^{*(2)}(\boldsymbol{\mu}_k)] \in \arg\min_{\boldsymbol{w}\in C_{12}} \big\{ &f^{(1)}(w^{(1)}) + f^{(2)}(w^{(2)}) \\ &- \mu_k^{(1)}(w^{(1)} + w^{(2)}) + \mu_k^{(2)} w^{(1)} + \mu_k^{(3)} w^{(2)} \big\}, \end{aligned} \tag{16}$$

$$z_k^{*(3)}(\boldsymbol{\mu}_k) \in \arg\min_{w\in[0,1]} \big\{ f^{(3)}(w) - \mu_k^{(2)} w \big\}, \tag{17}$$

$$z_k^{*(4)}(\boldsymbol{\mu}_k) \in \arg\min_{w\in[0,1]} \big\{ f^{(4)}(w) - \mu_k^{(3)} w \big\}, \tag{18}$$

$$\lambda_{k+1}^{(1)} = [\lambda_k^{(1)} + \alpha(\frac{1}{2} - z_k^{*(1)}(\boldsymbol{\mu}_k) - z_k^{*(2)}(\boldsymbol{\mu}_k))]^+, \tag{19}$$

$$\lambda_{k+1}^{(2)} = [\lambda_k^{(2)} + \alpha(z^{(1)} - z_k^{*(3)}(\boldsymbol{\mu}_k))]^+, \tag{20}$$

$$\lambda_{k+1}^{(3)} = [\lambda_k^{(3)} + \alpha(z^{(2)} - z_k^{*(4)}(\boldsymbol{\mu}_k))]^+, \tag{21}$$

where $C_{12} := \{\boldsymbol{z} \in \mathbb{R}^2_+ \mid \|\boldsymbol{z}\|_1 \leq 1\}$. When $\boldsymbol{\mu}_k = \boldsymbol{\lambda}_k$ then (19)-(21) are equivalent to the classical dual subgradient update. Since the objective function is separable and constraints are linear it can be seen that the optimisation can be solved in a distributed manner.

Note from Equations (16)-(18) and Figure 2 that we can compute the dual function without knowing the mean arrival rate at each node. However, updates (19)-(21) require us to know the mean packet arrival rates.

C. Optimisation with Unsynchronised Updates, Discrete Actions and Queue Occupancies as Approximate Multipliers

Suppose now that at the beginning of each slot a node decides to transmit a packet depending on a transmission policy. The transmission policy of a node at a given time t consists of a single action. The action set of node 1 is $\{[0, 0], [1, 0], [0, 1]\}$ where action $[1, 0]$ denotes transmitting a packet to node 2, $[0, 1]$ to node 3 and $[0, 0]$ to doing nothing. Action set of nodes 2 and 3 is $\{0, 1\}$ where 1 denotes transmitting a packet out of the system and 0 to doing nothing. The transmission policy of the system at a given time t is given by

$$X := \begin{bmatrix} 0&1&0&0&1&0&0&1&0&0&1&0 \\ 0&0&1&0&0&1&0&0&1&0&0&1 \\ 0&0&0&1&1&1&0&0&0&1&1&1 \\ 0&0&0&0&0&0&1&1&1&1&1&1 \end{bmatrix},$$

where the columns of matrix X contain the different permutations of possible transmission policies of the nodes in the system.

Packets arrive into the system (queue of node 1) at discrete times $t = 1, 2, 3, \dots$ and they satisfy that $|\sum_{i=1}^t (b_i - \frac{1}{2})| \leq 1$ where $b_i \in \{0, 1\}$ denotes receiving a packet or not. We use vector $\boldsymbol{Q}_t = [Q_t^{(1)}, Q_t^{(2)}, Q_t^{(3)}]^T$ to denote the queue occupancy at nodes 1, 2, 3 respectively.

Suppose the transmission policy of the system is updated at times $t = 1, 2, 3, \dots.$ and is obtained as shown in Equation (23) in [1],

$$\boldsymbol{x}_k \in \arg\min_{\boldsymbol{x}\in X} \Big\| \Big(\sum_{i=1}^{k} \boldsymbol{z}^*(\boldsymbol{\mu}_i) - \sum_{i=1}^{k-1} \boldsymbol{x}_i \Big) - \boldsymbol{x} \Big\|_\infty \tag{22}$$

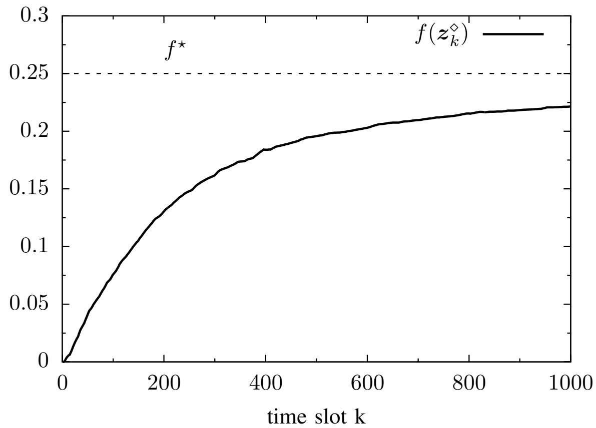


Fig. 3: Illustrating the converge of $f(\boldsymbol{z}_k^\diamond)$ to $f^\star = 0.25$ in the unsynchronised network example with step size $\alpha = 0.01$ and initial condition $\boldsymbol{Q} = \boldsymbol{0}$.

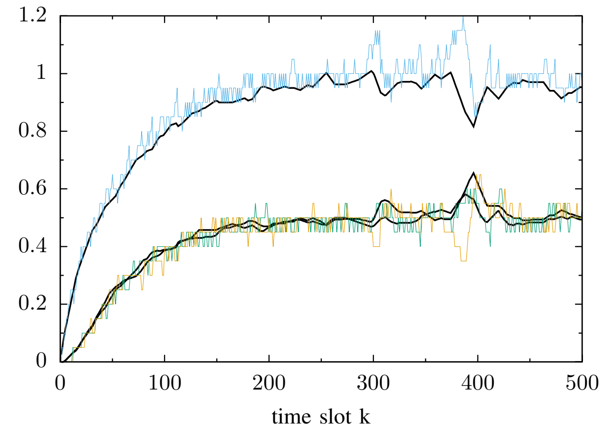


Fig. 4: Illustrating the convergence of the α-scaled queues to a ball around $\lambda^{\star(1)} = 1$, $\lambda^{\star(2)} = 0.5$ and $\lambda^{\star(3)} = 0.5$. Blue, green and orange lines correspond, respectively, to the α-scaled queues $Q^{(1)}$, $Q^{(2)}$ and $Q^{(3)}$ and black lines correspond to the associated Lagrange multipliers.

where a vector $\boldsymbol{x} \in X$ is a transmission policy of the system and $\boldsymbol{z}^*(\boldsymbol{\mu}_k) := [z^{*(1)}(\boldsymbol{\mu}_k), \dots, z^{*(4)}(\boldsymbol{\mu}_k)]^T$ is given by the subgradient updates

$$[z^{*(1)}(\boldsymbol{\mu}_k), z^{*(2)}(\boldsymbol{\mu}_k)] \in \arg\min_{\boldsymbol{w} \in C_{12}} \{f^{(1)}(w^{(1)}) + f^{(2)}(w^{(2)}) - \mu_k^{(1)}(w^{(1)} + w^{(2)}) + \mu_k^{(2)} w^{(1)} + \mu_k^{(3)} w^{(2)}\}, \quad (23)$$

$$z^{*(3)}(\boldsymbol{\mu}_k) \in \arg\min_{w \in [0,1]} \{f^{(3)}(w) - \mu_k^{(2)} w\}, \quad (24)$$

$$z^{*(4)}(\boldsymbol{\mu}_k) \in \arg\min_{w \in [0,1]} \{f^{(4)}(w) - \mu_k^{(3)} w\}, \quad (25)$$

where here $\boldsymbol{\mu}_k = \alpha \boldsymbol{Q}_k$ at each iterate with probability $1/5$, otherwise $\boldsymbol{\mu}_k = \boldsymbol{\mu}_{k-1}$, *i.e.*, updates are unsynchronised. Importantly, as shown in [1] update (22) keeps the scaled queue occupancy close to the Lagrange multipliers, and therefore Theorem 1 applies. Note as well from Equations (23)-(25) that as explained in the Section V-B the dual function can be computed without knowledge of the packet arrivals mean rates.

We simulate the network with objective functions $f^{(j)}(z^{(j)}) = (z^{(j)})^2$ for all $j = 1 \dots, 4$ with initial condition $\boldsymbol{Q} = \boldsymbol{0}$ and step size $\alpha = 0.01$. See in Figure 3 that as expected from Theorem 1 we have that $f(\boldsymbol{z}_k^\diamond)$ converges to $f^\star$. Figure 4 shows the scaled queue occupancy of the network and the Lagrange multipliers stay close and both converge to a ball around $\boldsymbol{\lambda}^\star$.

VI. Conclusions

We consider the subgradient method for the dual problem in convex optimisation with approximate multipliers, *i.e.*, the subgradient used in the update of the dual variables is obtained using an approximation of the true Lagrange multipliers. This problem is interesting for optimisation problems where the exact Lagrange multipliers might not be readily accessible. We show that the running average of the associated primal points can get arbitrarily close to the set of primal optima as step size α is reduced. Further, this running average is asymptotically attracted to a feasible point for every $\alpha > 0$. Applications of the analysis include unsynchronised subgradient updates in the dual variable update and unsynchronised max-weight scheduling.

References

[1] V. Valls and D. Leith, "On the relationship between queues and multipliers," in *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, Sept 2014, pp. 250–255.

[2] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*. Springer Berlin Heidelberg, 1985.

[3] B. Polyak, *Introduction to Optimization*. Optimization Software, Inc.,, 1987.

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

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

[6] A. Nedić and D. P. Bertsekas, "The effect of deterministic noise in subgradient methods," *Mathematical programming*, vol. 125, no. 1, pp. 75–99, 2010.

[7] K. Kiwiel and P. Lindberg, "Parallel subgradient methods for convex optimization," in *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, ser. Studies in Computational Mathematics. Elsevier, 2001, vol. 8, pp. 335 – 344.

[8] N. Gatsis and G. Giannakis, "Asynchronous subgradient methods with unbounded delays for communication networks," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, Dec 2012, pp. 5870–5875.

[9] A. Nedić and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1757–1780, 2009. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/070708111

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

[11] 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.

[12] 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.

[13] 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.

[14] V. Valls and D. J. Leith, "Max-weight revisited: Sequences of non-convex optimisations solving convex optimisations," *Networking, IEEE/ACM Transactions on*, in press.