Fifty-fourth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 27 - 30, 2016

Subgradient Methods with Perturbations in Network Problems

Víctor Valls, Douglas J. Leith
Trinity College Dublin

***Abstract*—We study the impact of perturbations on the convergence of the subgradient method for the dual problem in constrained convex optimisation. Perturbations are likely to be present in practical implementations of the subgradient method and can affect either the computation of a subgradient, the update of the dual variables, or both. In the context of networks, perturbations can be related to the exchange of network information, time varying channel conditions, discrete actions, *etc*. The analysis presented in this paper is general, and establishes the conditions under which the objective function will converge to the optimum asymptotically. With an example, we illustrate how the analysis can be applied to network flow problems where the intensity of the flows that arrive in the system changes over time.**

I. Introduction

Lagrange duality plays a prominent role in constrained convex optimisation and in modelling many interesting problems in networking. One of the appealing features of formulating the dual problem is that the dual variables (*i.e.* Lagrange multipliers) can be identified with physical or real quantities. For instance, in economics Lagrange multipliers can be identified with prices [1], in communication networks with scaled queue occupancies [2], and in electrical networks with potentials [3]. Identifying physical quantities with Lagrange multipliers can provide a useful insight in many practical problems [4], but can also allow some typically thought non-convex problems to be cast as convex optimisations [5].

In this paper, we study the impact of perturbations on the convergence of the subgradient method for the dual problem in constrained convex optimisation. One of the motivations for considering perturbations in the subgradient method is that in *practical* implementations the update of the dual variables (prices, queues, potentials), or the computation of a subgradient of the dual function, might be affected by noise or have errors. These perturbations can be used to encompass a wide range of important problems currently being discussed in the community, including asynchronous subgradient updates, distributed or parallel dual variables updates, discrete actions, actions with costs, *etc*. Perturbations can also be used to model important problems in networking, *e.g.* network resource allocation problems where the amount of resources that needs to be allocated changes or fluctuates over time. Note that this is actually the case in real communication networks since the load or capacity of the system changes depending on the number of users in the system, users' demand of bandwidth, channel conditions, *etc.*

Multiple perturbations can be present in the subgradient method, and they do not need to be independent. In fact, by correlating perturbations it is possible to model interesting problems such as expensive multipliers communication or expensive subgradient updates. For example, in communications networks the exchange of network state information (Lagrange multipliers) consumes resources (bandwidth), but at the same time controls the accuracy with which the optimisation problem can be solved. A similar problem appears in high-performance computing where a subgradient update (*i.e.* the selection of a new action or a change of configuration) consumes CPU resources, and so affects the amount of CPU time that can be allocated to tasks or jobs.

A. Related Work

Subgradient methods for solving nondifferentiable problems have been studied extensively under various step size rules by Polyak [6], Ermoliev [7] and Shor [8], or more recently by Bertsekas [9] and Nedić [10]. Approximate solutions to convex problems under an averaging scheme have been studied by Nedić in [11] and [12]. The work in [11] assumes that the dual function can be computed efficiently, and the work in [12] considers a sequence of primal-dual subgradient updates. See the related work in [11] for a good reference on primal averaging schemes. Perturbations regarding the inexact computation of the subgradient are not new and have been treated in previous work in terms of ϵ-subgradients [9], approximate Lagrange multipliers [13], or deterministic noise [14].

II. Preliminaries

A. Notation

Vectors and matrices are written, respectively, in lower and upper case, and all vectors are in column form. Since we usually work with sequences we will use subscript to indicate an element in a sequence, and parenthesis to indicate an element in a vector. For example, for a sequence $\{x_k\}$ of vectors from $\mathbb{R}^n$ we have that $x_k = [x_k(1), \dots, x_k(n)]^T$ where $x_k(j),\ j = 1, \dots, n$ is the j'th component of the k'th vector in the sequence. For two points $x, y \in \mathbb{R}^n$ we write $x \succ y$ when the $x(j) > y(j)$ for all $j = 1, \dots, n$, and $x \succeq y$ when $x(j) \geq y(j)$. Finally, we will use $[\cdot]^+$ to denote the projection of a vector $x \in \mathbb{R}^n$ onto the nonnegative orthant, *i.e.* $[x]^+ = [\max\{0, x(1)\}, \dots, \max\{0, x(n)\}]^T$.

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

B. Convex Optimisation

Consider the following convex optimisation problem P in standard form:

$$\begin{aligned} \underset{x\in X}{\text{minimise}} \quad & f(x) \\ \text{subject to} \quad & g_j(x) \le 0 \qquad j=1,\dots,m \end{aligned} \tag{1}$$

where $f, g_j : X \to \mathbb{R}$ are convex functions and X is a convex subset from $\mathbb{R}^n$. We will assume that set $X_0 := \{x \in X \mid g_j(x) \le 0,\ j=1,\dots,m\} \neq \emptyset$, and so problem P is feasible. Also, and using standard notation, we will define $f^\star := \min_{x\in X_0} f(x)$ and $x^\star \in \arg\min_{x\in X_0} f(x)$.

The Lagrangian of problem P is given by

$$L(x,\lambda) = f(x) + \lambda^T g(x),$$

where $g(x) = [g_1(x),\dots,g_m(x)]^T$ and $\lambda \in \mathbb{R}^m_+$, and the Lagrange dual function is

$$q(\lambda) := \inf_{x\in X} L(x,\lambda).$$

The following assumptions are key in our work.

Assumption 1 (Bounded Set)**.** *X is convex and bounded.*

Assumption 2 (Slater Condition)**.** $\mathrm{relint}(X_0)$ *is non-empty,* i.e. *there exists a point $x \in X$ such that $g_j(x) < 0$ for all $j = 1,\dots,m$.*

Assumption 1 ensures that the dual function is Lipschitz continuous, and Assumption 2 that strong duality holds. Namely, the solution of the Lagrange dual problem P^D,

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

coincides with the solution of the primal problem P. That is, $\max_{\lambda\succeq 0} q(\lambda) =: q(\lambda^\star) = f^\star$ where $\lambda^\star \in \arg\max_{\lambda\succeq 0} q(\lambda)$. Another consequence of Assumption 2 is that the set of dual optima is bounded.

The following lemma is a straightforward derivation of [11, Lemma 1].

Lemma 1 (Bounded Dual Set)**.** *Suppose Assumption 2 holds,* i.e. *there exists a point $\bar{x} \in X$ such that $g(\bar{x}) \prec 0$. Then,*

$$\lambda^\star(j) \le \frac{f(\bar{x}) - f^\star}{g_j(\bar{x})} \qquad j = 1,\dots,m.$$

Proof: Observe that since $\bar{x}$ is a feasible point we must have that $f^\star := \inf_{x\in X_0} f(x) \le f(\bar{x})$. Further, since $\lambda^\star \succeq 0$ and $g(\bar{x}) \prec 0$ then $\lambda^\star(j) g_j(\bar{x}) \le 0$ for all $j = 1,\dots,m$. Hence, we have that $\lambda^\star(j) g_j(\bar{x}) \le 0 \le f(\bar{x}) - f^\star$ and the stated result follows. ∎

One of the consequence of Lemma 1 is that there exists a constant $\lambda^\diamond \ge 0$ such that

$$\max_{\lambda\in\Lambda(\lambda^\diamond)} q(\lambda) = \max_{\lambda\in\mathbb{R}^m_+} q(\lambda) = f^\star, \tag{3}$$

where $\Lambda(\lambda^\diamond) := \{\lambda \in \mathbb{R}^m_+ \mid \lambda_j \le \lambda^\diamond,\ j = 1,\dots,m\}$ is a bounded set.

C. Classic Subgradient Method

Problem P^D is an unconstrained concave maximisation problem that can be solved using the subgradient method. In short, the subgradient method for the Lagrange dual problem consists of the following update:

$$\lambda_{k+1} = [\lambda_k + \alpha_k \partial q(\lambda_k)]^+, \tag{4}$$

where $\lambda_1 \in \mathbb{R}^m_+$, $\alpha_k > 0$ is a step size, and $\partial q(\lambda_k)$ is a subgradient of q at λ_k. In this work we will make use of a constant step size and so have $\alpha_k = \alpha$ for all k. Recall that a subgradient of q at λ_k is $g(x_k)$, where

$$x_k \in \arg\min_{x\in X} \ L(x,\lambda_k), \tag{5}$$

i.e. x_k is a solution of the unconstrained convex optimisation problem $\min_{x\in X} L(x,\lambda_k)$.

III. Optimisation with Perturbations

A. Perturbed Problem

Consider convex optimisation problem

$$\begin{aligned} \underset{x\in X}{\text{minimise}} \quad & f(x) \\ \text{subject to} \quad & g(x) \preceq 0 \end{aligned} \tag{6}$$

where $g : \mathbb{R}^n \to \mathbb{R}^m$, and the following δ*-perturbed* version of problem (6)

$$\begin{aligned} \underset{x\in X}{\text{minimise}} \quad & f(x) \\ \text{subject to} \quad & g(x) + \delta \preceq 0 \end{aligned} \tag{7}$$

where $\delta \in \mathbb{R}^m$ is an *unknown* perturbation in the constraints. Even though perturbation δ is not known we will assume that problem (7) is feasible, and that the Slater condition is satisfied, *i.e.* there exists a point $\bar{x} \in X$ such that $g(\bar{x}) + \delta \prec 0$. The difficulty in solving problem (7) is that the problem itself is *not known*, and so it is not possible to use standard optimisation algorithms (such as interior-point methods [15]) to solve it. This kind of problems are, however, not uncommon and appear in stochastic control, where controllers have to be designed without perfect knowledge of the randomness in the system. In communication networks sources of randomness can be, for example, packet arrivals or time varying channel conditions.

In the next section we show how to solve optimisation problem (7) by using perturbations in the updates of the dual variables of the subgradient method. Our key assumption is that we can observe a value δ_i at each iteration of the subgradient method, and that $\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \delta_i = \delta$. In order to emphasise that the problem that we aim to solve is the perturbed one, we define the *perturbed* Lagrangian

$$\tilde{L}(x,\lambda) := f(x) + \lambda^T (g(x) + \delta), \tag{8}$$

and the *perturbed* dual function

$$\tilde{q}(\lambda) := \inf_{x\in X} \tilde{L}(x,\lambda). \tag{9}$$

Also, and to avoid confusions, we will use $f^\star(\delta) = q(\lambda^\star(\delta))$ where $\lambda^\star(\delta) = \arg\max_{\lambda\succeq 0} \tilde{q}(\lambda)$ to denote a solution of the δ-perturbed problem.

B. Subgradient Method with Perturbations

In this section we study the subgradient method where the computation of a subgradient of the dual function (5) and the update of the dual variables (4) have perturbations.

1) Perturbations in the computation of a subgradient: First of all observe that

$$\begin{aligned} x_k &\in \arg\min_{x\in X} \tilde{L}(x,\lambda), \\ &= \arg\min_{x\in X}\{f(x)+\lambda_k^T(g(x)+\delta)\}, \\ &= \arg\min_{x\in X}\{f(x)+\lambda_k^T g(x)\}, \end{aligned}$$

and therefore x_k can be obtained irrespectively of the perturbation δ. We extend update (5) to allow perturbations or errors in the Lagrange multiplier, which will result in an "inexact" computation of the subgradient of the dual function. That is,

$$x_k \in \arg\min_{x\in X} \ \tilde{L}(x,\mu_k), \tag{10}$$

where $\mu_k = \lambda_k + \epsilon_k$ with $\epsilon_k \in \mathbb{R}^m$. We will usually refer to μ_k as an approximate or perturbed Lagrange multiplier, and it will capture the fact that in some optimisation problems the exact Lagrange multipliers are not known or have errors. For instance, in distributed optimisation a node might not have access to the exact Lagrange multipliers in the system due to transmission delays or losses, however, a delayed or outdate version of the *true* Lagrange multipliers might be available instead. Approximate Lagrange multipliers can also allow to capture asynchronous subgradient updates, and to use scaled queue occupancies as surrogates for the Lagrange multipliers.

2) Perturbations in the update of the dual variables: The other perturbation we consider is in the update of the dual variables, *i.e.* we have

$$\lambda_{k+1} = [\lambda_k + \alpha(g(x_k)+\delta_k)]^+, \tag{11}$$

where $\{\delta_k\}$ is a sequence of points from $\mathbb{R}^m$ such that $\lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^k \delta_i = \delta$. Sequence $\{\delta_k\}$ could be a realisation of a random variable, however, there is no requirement for the random variables to be independent and identically distributed (i.i.d.). A sequence $\{\delta_k\}$ can also be selected in each iteration in order to model a particular problem characteristic. For example, if λ_k represents a queue occupancy and $g(x_k)$ a packet transmission from the queue, a lossy link can be modelled using perturbations by selecting $\delta_k = -g(x_k)$. Namely, we enforce that $\lambda_{k+1} = \lambda_k$ and so the transmitted packet cannot "leave" the queue and must be transmitted again. Importantly, selecting $\delta_k = -g(x_k)$ for some k does not guarantee that the Slater condition is satisfied in the perturbed problem (7), neither that the perturbed problem is even feasible. Special care has to be taken to ensure that the Slater condition is satisfied when perturbations are chosen to model some kind of phenomena in the system.

C. Convergence

We start by presenting the following lemma.

Lemma 2 (Subgradient Method with Perturbations)**.** *Consider the setup of optimisation problem P^D and updates (10) and (11). Suppose $\{\delta_k\}$, $\{\epsilon_k\}$ are two sequence of points from $\mathbb{R}^m$ and that $\lambda_1 = 0$. Then,*

$$-\frac{1}{k}\sum_{i=1}^k \Gamma_i - \frac{\|\theta\|_2^2}{2\alpha k} \le \tilde{q}\left(\frac{1}{k}\sum_{i=1}^k \lambda_i\right) - \tilde{q}(\theta).$$

where $\Gamma_i = \frac{\alpha}{2}\|g(x_i)+\delta_i\|_2^2 + (\lambda_i-\theta)^T(\delta_i-\delta) + 2\|\epsilon_i\|_2\|g(x_i)+\delta\|_2$, and θ is any vector from $\mathbb{R}^m_+$.

Proof: Observe that for any vector $\theta\in\mathbb{R}^m$ we have that

$$\begin{aligned} \|\lambda_{k+1}-\theta\|_2^2 &= \|[\lambda_k+\alpha(g(x_k)+\delta_k)]^+ - \theta\|_2^2 \\ &\le \|\lambda_k+\alpha(g(x_k)+\delta_k)-\theta\|_2^2 \\ &= \|\lambda_k-\theta\|_2^2 + \alpha^2\|g(x_k)+\delta_k\|_2^2 \\ &\quad + 2\alpha(\lambda_k-\theta)^T(g(x_k)+\delta_k). \end{aligned}$$

We can write the last term in the RHS of the last equation as $(\lambda_k-\theta)^T(g(x_k)+\delta_k) = (\lambda_k-\theta)^T(g(x_k)+\delta) + (\lambda_k-\theta)^T(\delta_k-\delta)$ and obtain

$$\begin{aligned} &\|\lambda_{k+1}-\theta\|_2^2 - \|\lambda_k-\theta\|_2^2 \\ &\le \alpha^2\|g(x_k)+\delta_k\|_2^2 + 2\alpha(\lambda_k-\theta)^T(g(x_k)+\delta+(\delta_k-\delta)). \end{aligned}$$

Now, observe that since

$$x_k \in \arg\min_{x\in X} L(x,\lambda_k) = \arg\min_{x\in X}\tilde{L}(x,\lambda_k) \tag{12}$$

for all $\lambda_k\in\mathbb{R}^m_+$, we can write

$$\begin{aligned} &(\lambda_k-\theta)^T(g(x_k)+\delta) \\ &\quad = (\lambda_k-\theta)^T(g(x_k)+\delta) + f(x_k) - f(x_k) \\ &\quad = \tilde{L}(x_k,\lambda_k) - \tilde{L}(x_k,\theta) \\ &\quad \le \tilde{L}(x_k,\lambda_k) - \tilde{q}(\theta), \end{aligned}$$

where the last equation follows since $\tilde{q}(\theta) = \inf_{x\in X}\tilde{L}(x,\theta) \le \tilde{L}(x_k,\theta)$. Hence,

$$\begin{aligned} &\|\lambda_{k+1}-\theta\|_2^2 - \|\lambda_k-\theta\|_2^2 \\ &\quad \le \alpha^2\|g(x_k)+\delta_k\|_2^2 + 2\alpha(\lambda_k-\theta)^T(\delta_k-\delta) \\ &\qquad + 2\alpha(\tilde{L}(x_k,\lambda_k)-\tilde{q}(\theta)). \end{aligned}$$

Now, observe that since $\tilde{L}(x_k,\lambda_k) = \tilde{L}(x_k,\lambda_k) - \tilde{L}(x_k,\mu_k) + \tilde{L}(x_k,\mu_k) \le |\tilde{L}(x_k,\lambda_k)-\tilde{L}(x_k,\mu_k)| + \tilde{L}(x_k,\mu_k) = \tilde{q}(\mu_k) + |\tilde{L}(x_k,\lambda_k)-\tilde{L}(x_k,\mu_k)| = \tilde{q}(\mu_k) + |(\lambda_k-\mu_k)^T(g(x_k)+\delta)| = \tilde{q}(\mu_k) + |\epsilon_k^T(g(x_k)+\delta)| \le \tilde{q}(\mu_k) + \|\epsilon_k\|_2\|g(x_k)+\delta\|_2 = \tilde{q}(\mu_k) - \tilde{q}(\lambda_k) + \tilde{q}(\lambda_k) + \|\epsilon_k\|_2\|g(x_k)+\delta\|_2 \le |\tilde{q}(\mu_k)-\tilde{q}(\lambda_k)| + \tilde{q}(\lambda_k) + \|\epsilon_k\|_2\|g(x_k)+\delta\|_2 \le \tilde{q}(\lambda_k) + 2\|\epsilon_k\|_2\|g(x_k)+\delta\|_2$, we have that

$$\begin{aligned} &\|\lambda_{k+1}-\theta\|_2^2 - \|\lambda_k-\theta\|_2^2 \\ &\quad \le \alpha^2\|g(x_k)+\delta_k\|_2^2 + 2\alpha(\lambda_k-\theta)^T(\delta_k-\delta) \\ &\qquad + 2\alpha(\tilde{q}(\lambda_k)-\tilde{q}(\theta)) + 4\alpha\|\epsilon_k\|_2\|g(x_k)+\delta\|_2. \end{aligned}$$

Summing from $i=1,\ldots,k$ we obtain $\sum_{i=1}^k(\|\lambda_{i+1}-\theta\|_2^2 - \|\lambda_i-\theta\|_2^2) \le \sum_{i=1}^k(\alpha^2\|g(x_i)+\delta_i\|_2^2 + 2\alpha(\lambda_i-\theta)^T(\delta_i-\delta) +$

$2\alpha(\tilde{q}(\lambda_i) - \tilde{q}(\theta)) + 4\alpha\|\epsilon_i\|_2\|g(x_i) + \delta\|_2)$, and rearranging terms and dividing by $2\alpha k$ yields

$$-\frac{1}{k}\sum_{i=1}^{k}\Big(\frac{\alpha}{2}\|g(x_i)+\delta_i\|_2^2 + (\lambda_i-\theta)^T(\delta_i-\delta) + 2\|\epsilon_i\|_2\|g(x_i)+\delta\|_2\Big) + \frac{\|\lambda_{k+1}-\theta\|_2^2 - \|\lambda_1-\theta\|_2^2}{2\alpha k} \leq \frac{1}{k}\sum_{i=1}^{k}\tilde{q}(\lambda_i) - \tilde{q}(\theta).$$

Finally, by the convexity of $-q$ we can write

$$\frac{1}{k}\sum_{i=1}^{k}\tilde{q}(\lambda_i) \leq \tilde{q}\left(\frac{1}{k}\sum_{i=1}^{k}\lambda_i\right), \tag{13}$$

and the stated result follows. ∎

Lemma 2 is stated in a general form and provides a lower bound on the difference $\tilde{q}(\frac{1}{k}\sum_{i=1}^{k}\lambda_i) - \tilde{q}(\theta)$, where θ is any vector from $\mathbb{R}_+^m$. The assumption that $\lambda_1 = 0$ is not essential and we just make it to obtain a simpler bound. Now, let $\theta = \bar{\delta}_k := \frac{1}{k}\sum_{i=1}^{k}\delta_i$ and suppose that there exists a point $x \in X$ such that $g(x)+\bar{\delta}_k \prec 0$, that is, the Slater condition is satisfied in problem (7) with perturbation $\bar{\delta}_k$ (instead of δ). Then,

$$-\frac{1}{k}\sum_{i=1}^{k}\Gamma_i - \frac{\|\lambda^\star(\bar{\delta}_k)\|_2^2}{2\alpha k} \leq \tilde{q}\left(\frac{1}{k}\sum_{i=1}^{k}\lambda_i\right) - \tilde{q}(\lambda^\star(\bar{\delta}_k)) \leq 0,$$

where the upper bound follows directly from the fact that $\tilde{q}(\lambda^\star(\bar{\delta}_k)) \geq \tilde{q}(\lambda)$ for all $\lambda \in \mathbb{R}_+^m$. Note that since $\lambda^\star(\bar{\delta}_k)$ is bounded by Lemma 1, then, when $k \to \infty$ we have that $-\bar{\Gamma} \leq \tilde{q}(\frac{1}{k}\sum_{i=1}^{k}\lambda_i) - \tilde{q}(\lambda^\star(\bar{\delta}_k)) \leq 0$ where $\bar{\Gamma} := \lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}\Gamma_i$. That is, in order for the bound to be useful we need that $\bar{\Gamma}$ is bounded and small in appropriate sense, which will depend on the assumptions made on sequences $\{\epsilon_k\}$ and $\{\delta_k\}$.

Before making any assumptions on the perturbations we present the following theorem, which establishes a bound on how approximate primal solutions can be recovered from the perturbed subgradient method.

Theorem 1 (Approximate Primal Solutions). *Consider the setup of Lemma 2 and let $\theta = \lambda^\star(\bar{\delta}_k)$ where $\lambda^\star(\bar{\delta}_k) \in \arg\max_{\lambda\succeq 0}\{\tilde{q}(\lambda)\}$ and $\bar{\delta}_k = \frac{1}{k}\sum_{i=1}^{k}\delta_i$. If there exists a point $x \in X$ such that $g(x) + \bar{\delta}_k \prec 0$ the following bound holds*

$$\begin{aligned}
&-\frac{1}{k}\sum_{i=1}^{k}\Gamma_i - \frac{\|\lambda^\star(\bar{\delta}_k)\|_2^2}{2\alpha k} - \left(\frac{\lambda_{k+1}}{\alpha k}\right)^T\left(\frac{1}{k}\sum_{i=1}^{k}\lambda_i\right)\\
&\leq f(\bar{x}_k) - f^\star(\bar{\delta}_k)\\
&\leq \frac{1}{k}\sum_{i=1}^{k}(\lambda_i^T(\bar{\delta}_k - \delta_i) + \frac{\alpha}{2}\|g(x_i)+\delta_i\|_2^2\\
&\qquad + \|\epsilon_i\|_2\|g(x_i)+\delta_i\|_2),
\end{aligned}$$

where $f^\star(\bar{\delta}_k) = q(\lambda^\star(\bar{\delta}_k))$, $\Gamma_i = \frac{\alpha}{2}\|g(x_i)+\delta_i\|_2^2 + (\lambda_i - \lambda^\star(\bar{\delta}_k))^T(\delta_i - \delta) + 2\|\epsilon_i\|_2\|g(x_i)+\delta\|_2$, $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k}x_i$, and x_i is the primal variable obtained with update (10).

Proof: Consider the perturbed problem (7) with $\bar{\delta}_k$ and observe that

$$\frac{1}{k}\sum_{i=1}^{k}\tilde{L}(x_i,\mu_i) = \frac{1}{k}\sum_{i=1}^{k}\tilde{q}(\mu_i) \leq \tilde{q}\left(\frac{1}{k}\sum_{i=1}^{k}\mu_i\right) \leq \tilde{q}(\lambda^\star(\bar{\delta}_k)).$$

Further, since $\frac{1}{k}\sum_{i=1}^{k}\tilde{L}(x_i,\mu_i) = \frac{1}{k}\sum_{i=1}^{k}f(x_i) + \mu_i^T(g(x_i)+\bar{\delta}_k) \geq f(\bar{x}_k) + \frac{1}{k}\sum_{i=1}^{k}\mu_i^T(g(x_i)+\bar{\delta}_k)$, and $\mu_k = \lambda_k + \epsilon_k$ we have that

$$f(\bar{x}_k) - f^\star(\bar{\delta}_k) \leq -\frac{1}{k}\sum_{i=1}^{k}(\lambda_i+\epsilon_i)^T(g(x_i)+\bar{\delta}_k) \tag{14}$$

From Lemma 2 we have there exists a sequence $\{\beta_k\}$ of points from $\mathbb{R}_+ \cup \{+\infty\}$ such that $-\beta_k \leq \tilde{q}(\bar{\lambda}_k) - \tilde{q}(\lambda^\star(\bar{\delta}_k))$ for all k where $\bar{\lambda}_k = \frac{1}{k}\sum_{i=1}^{k}\lambda_i$. Since $\tilde{q}(\bar{\lambda}_k) \leq \tilde{L}(\bar{x}_k,\bar{\lambda}_k) = f(\bar{x}_k) + \bar{\lambda}_k^T(g(\bar{x}_k)+\bar{\delta}_k)$, by rearranging terms we have

$$-\beta_k - \left(\frac{1}{k}\sum_{i=1}^{k}\lambda_i\right)^T(g(\bar{x}_k)+\bar{\delta}_k) \leq f(\bar{x}_k) - f^\star(\bar{\delta}_k) \tag{15}$$

We now proceed to upper bound the RHS of (14). Observe that for *any* sequence $\{x_k\}$ in X we can write

$$\begin{aligned}
\|\lambda_{k+1}\|_2^2 &= \|[\lambda_k + \alpha(g(x_k)+\delta_k)]^+\|_2^2\\
&\leq \|\lambda_k + \alpha(g(x_k)+\delta_k)\|_2^2\\
&\leq \|\lambda_k\|_2^2 + \alpha^2\|g(x_k)+\delta_k\|_2^2 + 2\alpha\lambda_k^T(g(x_k)+\delta_k).
\end{aligned}$$

Rearranging terms and summing from $i = 1,\ldots,k$ we have that $\sum_{i=1}^{k}(\|\lambda_{i+1}\|_2^2 - \|\lambda_i\|_2^2) \leq \alpha^2\sum_{i=1}^{k}\|g(x_i)+\delta_i\|_2^2 + 2\alpha\sum_{i=1}^{k}\lambda_i^T(g(x_i)+\delta_i)$, and further rearranging and dividing by $2\alpha k$ yields

$$-\frac{1}{k}\sum_{i=1}^{k}\lambda_i^T(g(x_i)+\delta_i) \leq \frac{\alpha}{2k}\sum_{i=1}^{k}\|g(x_i)+\delta_i\|_2^2 + \frac{\|\lambda_1\|_2^2}{2\alpha k}.$$

Next, observe that since $\lambda_i^T(g(x_i)+\delta_i) = \lambda_i^T(g(x_i)+\delta_i+\bar{\delta}_k-\bar{\delta}_k)$, and $\lambda_1 = 0$ (by assumption) we can write

$$-\frac{1}{k}\sum_{i=1}^{k}\lambda_i^T(g(x_i)+\bar{\delta}_k) \leq \frac{1}{k}\sum_{i=1}^{k}\lambda_i^T(\bar{\delta}_k-\delta_i) + \frac{\alpha}{2k}\sum_{i=1}^{k}\|g(x_i)+\delta_i\|_2^2.$$

Finally, since by Cauchy-Schwarz $-\epsilon_i^T(g(x_i)+\delta_i) \leq \|\epsilon_i\|_2\|g(x_i)+\delta_i\|_2$ the upper bound now follows.

For the lower bound observe that $\lambda_{k+1} \succeq [\lambda_k + \alpha(g(x_k)+\delta_k)]^+ \succeq \lambda_k + \alpha(g(x_k)+\delta_k)$ and therefore $\alpha(g(x_k)+\delta_k) \preceq \lambda_{k+1} - \lambda_k$. Summing from $i = 1,\ldots,k$ we have $\alpha\sum_{i=1}^{k}(g(x_i)+\delta_i) \preceq \sum_{i=1}^{k}(\lambda_{i+1}-\lambda_i) \preceq (\lambda_{k+1}-\lambda_1)$. Dividing by αk yields

$$g(\bar{x}_k) + \bar{\delta}_k \preceq \frac{1}{\alpha k}\sum_{i=1}^{k}(g(x_i)+\delta_i) \preceq \frac{\lambda_{k+1}-\lambda_1}{\alpha k}.$$

Finally, multiplying both sides by $(\frac{1}{k}\sum_{i=1}^{k}\lambda_i)$ we have

$$\left(\frac{1}{k}\sum_{i=1}^{k}\lambda_i\right)^T (g(\bar{x}_k)+\bar{\delta}_k) \leq \left(\frac{\lambda_{k+1}}{\alpha k}\right)^T \left(\frac{1}{k}\sum_{i=1}^{k}\lambda_i\right),$$

which concludes the proof. ∎

Theorem 1 says that $f(\bar{x}_k)$ approaches $f^\star(\bar{\delta}_k)$ when the Slater condition is satisfied with perturbation $\bar{\delta}_k = \frac{1}{k}\sum_{i=1}^{k}\delta_i$. However, and as in Lemma 2, the bounds are loose because we have not made any assumptions on sequences $\{\epsilon_k\}$ and $\{\delta_k\}$. Next, we present two corollaries to Theorem 1 in which we make assumptions on sequences $\{\epsilon_k\}$ and $\{\delta_k\}$.

Corollary 1. *Consider the setup of Theorem 1, and suppose that $\|\epsilon_k\|_2 \leq \epsilon$ and $\delta_k = \delta$ for all k with $\epsilon > 0$ and $\delta \in \mathbb{R}^m$. Suppose there exists a point $x \in X$ such that $g(x) + \delta \prec 0$, and that Assumption 1 holds,* i.e. $\max_{x\in X}\|g(x)+\delta\|_2 := \sigma_g$ *is finite. Then, when $k \to \infty$ we have that*

$$|f(\bar{x}_k) - f^\star(\delta)| \leq \frac{\alpha}{2}\sigma_g^2 + 2\epsilon\sigma_g$$
$$\lim_{k\to\infty} g(\bar{x}_k) + \delta \preceq 0$$

Corollary 1 considers the inexact computation of a subgradient without perturbation δ_k on the update of the dual variables. Observe that term $\sigma_g^2/2$ can be made arbitrarily small by selecting α small. However, since term $2\epsilon\sigma_g$ does not depend on α we have that $f(\bar{x}_k)$ converges asymptotically to a ball around $f^\star(\delta)$.

The following corollary extends the previous one to consider perturbations δ_k.

Corollary 2. *Consider the setup of Theorem 1, and suppose $\{\delta_k\}$ is an ergodic stochastic process with $\mathbb{E}(\delta_k) = \delta$. Suppose $\mathbb{E}(\|\delta_k - \delta\|_2^2) = \sigma_\delta^2$, and that $\|\epsilon_k\|_2 \leq \epsilon$ for all k for some $\epsilon > 0$. Further, suppose that Assumption 1 holds,* i.e. $\max_{x\in X}\|g(x)+\delta\|_2 := \sigma_g$ *is finite. Then, when $k\to\infty$ we have that*

$$|\mathbb{E}(f(\bar{x}_k) - f^\star(\delta))| \leq \frac{\alpha}{2}(\sigma_g^2 + \sigma_\delta^2) + 2\epsilon\sigma_g$$
$$\lim_{k\to\infty} g(\bar{x}_k) + \delta \preceq 0$$

By considering perturbation δ_k we have added the term $\alpha(\sigma_\delta^2/2)$ to the bound in Corollary 1, which is related to the variance of the stochastic process $\{\delta_k\}$. Note that if we let $\sigma_\delta^2 = 0$ we then recover the bound in Corollary 1. Further, see that unlike with perturbations ϵ_k, with perturbations δ_k the effect of the perturbations can be reduced by selecting α small. Another interesting observation is that since σ_δ^2 has to be finite, then the stochastic process $\{\delta_k\}$ cannot be heavy-tailed.

IV. Numerical Example

In this section, we show how the analysis can be used in network flow problems where the intensity of the flows changes over time—that is, capture the fact that the flows in a network may depend on factors such as the users behaviour.

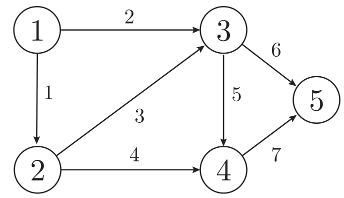

Figure 1: Illustrating the network of the numerical example in Section IV.

A. Problem Setup

Consider the network illustrated in Figure 1 with $m = 5$ nodes and $n = 7$ links.[1] Flows arrive in the system at nodes 1 and 2, and they must be allocated to links in order to reach node 5, where they will leave the system. The incidence matrix of the network is given by

$$A = \begin{bmatrix} -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

and we will assume that the links have unit capacity. The time in the system is divided in slots of equal duration, and at each time slot a node can decide whether to allocate a flow to a link. The goal of the problem is to design a distributed scheduling policy that minimises the cost of allocating flows to links—for example, suppose that a network operator charges for a link usage.

The convex formulation of the problem is

$$\begin{array}{ll} \underset{x\in X}{\text{minimise}} & f(x) = \sum_{j=1}^{n} f_j(x(j)) \\ \text{subject to} & Ax + b \preceq 0 \end{array} \tag{16}$$

where $X := \prod_{j=1}^{n} X_j$, $X_j := [0,1]$ for all $j = 1,\ldots,n$, $b \in \mathbf{R}^m$ is a vector containing the mean arrival/departure of flows in the network at each of the nodes, and $f_j : \mathbf{R} \to \mathbf{R}$ are convex functions that capture the cost of using each link $j = 1,\ldots,n$. Optimisation (16) can be solved with the dual subgradient method, and since the objective function is fully separable and constraints are linear, the computation of a (dual) subgradient can be carried out in a distributed manner. In particular, in each iteration we have updates

$$x_k(j) \in \arg\min_{x\in X_j}\{f_j(x) + \alpha r_k(j)x\}, \qquad j = 1,\ldots,n \tag{17}$$

$$\lambda_{k+1} = [\lambda_k + \alpha(Ax_k + B_k)]^+, \tag{18}$$

where $r_k = \lambda_k^T A$, and $B_k \in \mathbf{R}^m$ is a random variable that captures the intensity of the flows that arrive/leave the system in each of the nodes. Note that we have assumed that in each update (17) a node has perfect knowledge of the Lagrange multipliers in the system, however, this could be relaxed by using an approximate Lagrange multiplier μ_k in the update.

[1] The network is taken from the flow example in [16].

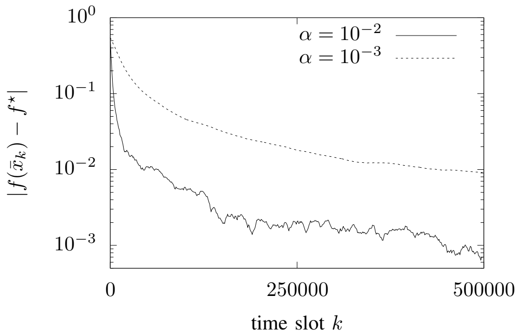

Figure 2: Illustrating the convergence of the utility function.

Namely, with μ_k we could capture imperfect network state information when allocating flows to links.

B. Simulation

We run updates (17) and (18) with $f_j(x(j)) = x(j)^2$ for every link $j = 1, \ldots, n$, $b = [0.2, 0.6, 0, 0, -1]^T$, $B_k(j)$, $j = 1, 2$ are Bernoulli with $\mathbb{E}(B_k(j)) = b(j)$, and $B_k(j)$ for $j = 3, 4, 5$ are equal to $b(j)$ for all k, *i.e.* nodes 3 and 4 do not contribute to changing the flow load in the system, and the service of node 5 is deterministic.

Figure 2 shows the convergence of $f(\bar{x}_k)$ to a ball around the optimum for $\alpha = \{10^{-2}, 10^{-3}\}$. Observe from Figure 2 that despite the asymptotic convergence established in Corollary 2, we have that $f(\bar{x}_k)$ is attracted to $f^\star$ for finite k. Note as well from the figure that with $\alpha = 10^{-2}$ we have better performance than with $\alpha = 10^{-3}$ for finite k. Nonetheless, from Corollary 2 we have that asymptotically, by using a smaller step size, we will recover a better solution.

V. Conclusions

We have studied the impact of perturbations on the convergence of the subgradient method for the dual problem in constrained convex optimisation. The study of perturbations is motivated because in practical implementations the dual variables updates and the computation of the subgradients of the dual function can be affected by noise or errors. Our results establish the asymptotic convergence of the objective function when the perturbations that affect the computation of a dual subgradient are bounded, and when the perturbations in the update of the dual variables are ergodic and have finite variance. With an example, we have shown how the analysis can be used in network flow problems where the intensity of the flows that arrive in the system changes over time.

References

[1] D. K. H. T. F. P. Kelly, A. K. Maulloo, “Rate control for communication networks: Shadow prices, proportional fairness and stability,” *The Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[2] V. Valls and D. J. Leith, “Descent With Approximate Multipliers is Enough: Generalising Max-Weight,” *ArXiv e-prints*, Nov. 2015.

[3] R. Rockafellar, *Network Flows and Monotropic Optimization*. John Wiley & Sons, Inc., Halsted Press, 1984.

[4] L. Huang and M. J. Neely, “Delay reduction via lagrange multipliers in stochastic network optimization,” *IEEE Transactions on Automatic Control*, vol. 56, no. 4, pp. 842–857, April 2011.

[5] V. Valls and D. J. Leith, “Max-weight revisited: Sequences of nonconvex optimizations solving convex optimizations,” *IEEE/ACM Transactions on Networking*, vol. in press, 2015.

[6] B. T. Polyak, “A general method of solving extremum problems,” *Soviet Mathematics Doklady*, vol. 8, no. 3, pp. 593–597, 1967.

[7] Y. M. Ermoliev, “Methods of solution of nonlinear extremal problems,” *Cybernetics*, vol. 2, no. 4, pp. 1–14, 1966.

[8] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*. Springer Science & Business Media, 2012, vol. 3.

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

[10] A. Nedić, “Subgradient methods for convex minimization,” Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[11] 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.

[12] 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.

[13] V. Valls and D. J. Leith, “Dual subgradient methods using approximate multipliers,” in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2015, pp. 1016–1021.

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

[15] Y. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

[16] S. Boyd, L. Xiao, A. Mutapcic, and J. Mattingley, “Notes on decomposition methods,” pp. 1–36, 2007.