On the Convergence of Limited Feedback Gradient Descent

Sindri

I. Introduction

A. Related Literature

[1], [2]

[3]

[4] studies the convergence of standard interference where the base station functions for power control in cellular wireless systems. the convergence for finding the

[5] considers quantized incremental gradient gradient over a network, where each node Unlike our work, in [5] quantized version of the primal variable but in the gradient step full gradient information is assumed.

[6], [7]

II. Problem Statement and Motivational Applications

In this paper we consider general problems on the form

$$\underset{\mathbf{p}\in\mathbb{R}^R}{\text{minimize}} \quad D(\mathbf{p}). \tag{1}$$

We denote by $D^\star$ and $\mathcal{P}^\star$ the optimal value and the set of optimizers to Problem (1), respectively. We use the notation D and $\mathbf{p}$ to highlight the applications of the developed theory to dual or primal decomposition [6] where (1) is a master problem and $\mathbf{p}$ represents prices or dual variables. Throughout this paper we make the following which widely hold in practice (see the following subsections).

Assumption 1. *$\mathcal{P}^\star$ is non-empty. Moreover, D is convex and continuously differentiable with L-continuous gradient.*

A. The Gradient Descent

It is well known that under Assumption 1 the gradient descent method given by the iterations

$$\mathbf{x}(t+1) = \mathbf{x}(t) - \gamma(t)\nabla f(\mathbf{x}(t)) \tag{2}$$

converges to $\mathcal{X}^\star$ or to a arbitarly small neigborhod of $\mathcal{X}^\star$ under varis step-size rules. For example, if either (a) the step-size $\gamma(t) = \gamma$ is fixed and $\gamma \in]0, 2/L]$, or (b) $\gamma(t)$ are non-summable but square-summable. If only the gradient direction is known, then recursion (2) reduce to

$$\mathbf{x}(t+1) = \mathbf{x}(t) - \gamma(t)\frac{\nabla f(\mathbf{x}(t))}{||\nabla f(\mathbf{x}(t))||}, \tag{3}$$

where (a) for fixed step-size $\gamma(t) = \gamma$ and arbitrary $\epsilon>0$ the step-size γ can be chosen small enough so that $f(\mathbf{x}(t)) - f^\star$ for all sufficiently large t and (b) if $\gamma(t)$ is non-summable but not square-summable the iterates $\mathbf{x}(t)$ converge to $\mathcal{X}^\star$.

We now provide some practical application examples.

B. Resources Allocation with Multiple Distributed Resources

Consider a network consisting of R and N resources/sources and users/destinations, respectively. Each source Then is to solve a

$$\begin{aligned} \underset{\mathbf{q}\in\mathbb{R}^{NR}}{\text{maximize}} \quad & \sum_{i=1}^{N} U_i(\mathbf{q}_i) - \sum_{j=1}^{R} C_j(s_j) \\ \text{subject to} \quad & \mathbf{q}_i \in \mathcal{Q}_i, \quad \text{for } i = 1, \cdots, N \\ & s_j \in \mathcal{S}_i, \quad \text{for } j = 1, \cdots, R \\ & \sum_{i=1}^{N} \mathbf{q}_i = \mathbf{s}, \end{aligned} \tag{RA}$$

where $\mathbf{q} = (\mathbf{q}_1, \cdots, \mathbf{q}_N) \in \mathbb{R}^{N\times R}$ and $\mathbf{s} = (s_1, \cdots, s_R) \in \mathbb{R}^R$. The Problem (RA) is coupled between the users due to the coupling constraint $\sum_{i=1}^{N} \mathbf{q}_i = \mathbf{s}$. A standard trick to decouple a problem on the form of (RA) is to consider the its dual problem related to the coupling constraint. The dual problem is on the form (1) where the dual function $D : \mathbb{R}^R \to \mathbb{R}$ is given by

$$D(\mathbf{p}) = \max_{q\in\mathcal{I}} L(\mathbf{q}, \mathbf{s}, \mathbf{p}) = L(\mathbf{q}(\mathbf{p}), \mathbf{s}(\mathbf{p}), \mathbf{p}), \tag{4}$$

where

$$L(\mathbf{q}, \mathbf{s}, \mathbf{p}) = \sum_{i=1}^{N} U_i(q_i) - \sum_{j=1}^{R} C_j(s_j) - \mathbf{p}^{\mathrm{T}}\left(\mathbf{s} - \sum_{i=1}^{N} \mathbf{q}_i\right),$$

and for all $i = 1, \cdots, N$ and $i = 1, \cdots, R$ we have

$$\mathbf{q}_i(\mathbf{p}) = \underset{\mathbf{q}_i\in\mathcal{Q}_i}{\text{argmax}}\ U_i(\mathbf{q}_i) - \mathbf{p}^{\mathrm{T}}\mathbf{q}_i, \tag{5}$$

$$s_j(p_j) = \underset{s_j\in\mathcal{S}_j}{\text{argmax}}\ -C_i(s_i) - p_j s_j. \tag{6}$$

We have the following result:

Lemma 1. *Suppose for all $i = 1, \cdots, N$ and $j = 1, \cdots, R$ there exists $\mu > 0$ such that U_i are μ-concave and C_j are μ-convex and $\mathcal{Q}_i$ and $\mathcal{S}_i$ are compact sets. Then, D is continuously differentiable on $\mathbb{R}^R$ and the gradient ∇D is $(R + L)/\mu$-Lipschitz continuous.*

Proof. The fact that D is continuously differentiable under the given assumptions follows directly from [8, Proposition 6.1.1], since $L(\mathbf{q}, \mathbf{s}, \mathbf{p})$ is strictly concave for fixed $\mathbf{p}$. Let us now show that ∇D is $(R + L)/\mu$-Lipschitz continuous. □

Due to Lemma 1 the Algorithms (**??**) and (**??**) when the gradient is broadcasted to the users However, as motivated in the introduction, in many applications there are communication constraints on the brodcasted message from controller to users. The results of this paper show that similar convergence

results as for (**??**) can be obtained even the gradient message is qantified to $\log_2(R+1)$ bits per iteration if the sources cooperate and to $\log_2(2N)$ if the do not cooperate.

C. DC Optimal Power Flow

D. Primal Decomposition

The applications mentioned above exploit dual decomposition. However, the gradient methods for solving a problem of the form (**??**) are used in primal decomposition. Hence the theory developed in this paper can also be applied to all the related applications. More details on primal decomposition and their applications are found here [6] where (1).

Clearly, there is a new theory needed to address all the complications mentioned above. In the following two sections we demonstrate such theory.

III. Quantized Gradient Descent Methods

As demonstrated in the previous sections, many applications of primal/dual decomposition demand economic communication where the broadcasted gradients are quantized to as few bits as possible. This motivates the current work, where we investigate general quantized gradient methods of the form

$$\mathbf{p}(t+1) = \mathbf{p}(t) - \gamma(t)\mathbf{d}(t), \tag{7}$$

with $\mathbf{d}(t) \in \mathcal{D}$, where $\mathcal{D}$ is a finite set of quantized gradient direction. We investigate what properties the quantization set $\mathcal{D}$ needs to have so that the recursion (7) can "solve" any problem of the form (1). To formally assert the meaning of such a proper quantization we make the following definition.

Definition 1. *We say that the set $\mathcal{D}$ is a* proper quantization *of the gradient directions if for every optimization problem of the form* (1) *where Assumption 1 holds and any initiation $\mathbf{p}(0)$ we can choose $\mathbf{d}(t) \in \mathbb{R}^R$ and $\gamma(t) \in \mathbb{R}_+$, for all $t \in \mathbb{N}$, in the recursion* (7) *such that*

$$\lim_{t\to\infty} dist(\mathbf{p}(t), \mathcal{P}^\star) = 0 \tag{8}$$

In other words, every limit point of $\mathbf{p}(t)$ is in $\mathcal{P}^\star$.

Definition 1 is not constructive since it's validation requires testing the dynamics (7) on every problem of from (1) where Assumption 1 holds. Neither does Definition 1 provide an algorithm from (7) since it does not say anything about how to choose $\mathbf{d}(t)$ when $\mathcal{D}$ is known to be a *proper quantization*. Moreover, Definition 1 gives no insight into other interesting properties of the quantization set $\mathcal{D}$; for example the minimal size of $|\mathcal{D}|$ which ensures a proper quantization or how the structure of $\mathcal{D}$ affects the convergence behavior of potential quantized gradient methods. All the issues mentioned above are addressed in this paper, particularly, answers to the following questions are provided:

A) What are equivalent constructive conditions to the set $\mathcal{D}$ being *proper quantization* (Definition 1) which can be used to easily determine if $\mathcal{D}$ *proper quantization* or to construct such sets?
B) Given a proper quantification $\mathcal{D}$, how can we construct an algorithm from (7) such that $\lim_{t\to\infty} \mathtt{dist}(\mathbf{p}(t), \mathcal{P}^\star) = 0$, i.e., choose the proper $d(t) \in \mathcal{D}$? Moreover, what are the number of bits needed to perform such an algorithm?
C) What is the connection between the fineness of the quntizeion, i.e., the size of $|\mathcal{D}|$, to the possible convergence of the algorithm?
D) What is the minimal quantization, i.e., size $|\mathcal{D}|$, so that $\mathcal{D}$ is *proper quantization*?

In the following subsections we answer each of the questions above, but refer to later sections for many of technical details.

A. Answer to Question A)

We now provide condition that are equivalent to $\mathcal{D}$ being a proper quantization (Definition 1) but are *constructive* and can be used to determine if a set $\mathcal{D}$ is proper quantization and to construct such $\mathcal{D}$. We then use this condition to provide number of proper quantization sets $\mathcal{D}$.

Definition 2. *We say that the set $\mathcal{D}$ is a θ-cover of the directions in $\mathbb{R}^N$ if $\theta > 0$ and for every $\mathbf{x} \in \mathbb{R}^N$ there exists $\mathbf{d} \in \mathcal{D}$ such that*

$$\cos(ang(\mathbf{x}, \mathbf{d})) \geq \theta. \tag{9}$$

We say that the θ-cover $\mathcal{D}$ is tight if there exists a vector $\mathbf{x} \in \mathcal{S}^R$ such that $\max_{\mathbf{d}\in\mathcal{D}} \cos(ang(\mathbf{x}, \mathbf{d})) = \theta$.

The following result asserts the equivalence between Definitions 1 and 2.

Proposition 1. *Consider a quantizatin set $\mathcal{D}$. $\mathcal{D}$ is a proper quantization (Definition 1) if and only if there exists $\theta > 0$ such that $\mathcal{D}$ is a θ-cover for some $\theta > 0$ (Definition 2).*

Proof. Let us start by showing by using contradiction, that $\mathcal{D}$ being a proper quantization implies that there exists $\theta > 0$ such that $\mathcal{D}$ is θ-cover. Suppose there does not exist $\theta > 0$ such that $\mathcal{D}$ is a θ-cover. Then, since $\mathcal{D}$ is finite, there exits $\mathbf{a} \in \mathbb{R}^R$ such that $\cos(\mathtt{ang}(\mathbf{a}, \mathbf{d})) \leq 0$ for all $\mathbf{d} \in \mathcal{D}$. In particular, we have for all $\mathbf{d} \in \mathcal{D}$ that

$$\langle \mathbf{a}, \mathbf{d} \rangle = \|\mathbf{a}\|\|\mathbf{d}\| \cos(\mathtt{ang}(\mathbf{a}, \mathbf{d})) \leq 0. \tag{10}$$

Therefore, it follows from Lemma 2, that if we choose $\mathbf{p}(0) = \mathbf{a}/||\mathbf{a}||$ then no matter how $\mathbf{d}(t) \in \mathcal{D}$ and $\gamma(t) \in \mathbb{R}^R$ are chosen, $||\mathbf{p}(t)|| \geq 1$ for all $t \in \mathbb{N}$. In particular, if we choose the optimization problem

$$\underset{\mathbf{p}}{\text{minimize}} \quad ||\mathbf{p}||, \tag{11}$$

which clearly has the unique optimal solution $\mathbf{p}^\star = \mathbf{0}$, then

$$\mathtt{dist}(\mathbf{p}(t), \mathcal{P}^\star) = ||\mathbf{p}(t)|| \geq 1, \tag{12}$$

for all $t \in \mathbb{N}$. Since (12) holds for all chooses $\mathbf{d}(t) \in \mathcal{D}$ and $\gamma(t) \in \mathbb{R}^R$, we can conclude that $\mathcal{D}$ is not a proper quantization.

To show that $\mathcal{D}$ being a θ-cover implies that $\mathcal{D}$ is a proper quantization, we show that for any problem of the form (1) where Assumption 1 we can choose of $\mathbf{d}(t) \in \mathcal{D}$ and $\gamma \in \mathbb{R}_+$ such that $\lim_{t\to\infty} \mathtt{dist}(\mathbf{p}(t), \mathcal{P}^\star) = 0$. The result follows directly from Proposition **??** in Section **??**.

□

Lemma 2. *Suppose there exists $\mathbf{a} \in \mathbb{R}^N$ such that $\langle \mathbf{a}, \mathbf{d} \rangle \leq 0$ for all $\mathbf{d} \in \mathcal{D}$. Then for any $\kappa \in \mathbb{R}_+$, there exists $\mathbf{x}(0) \in \mathbb{R}^N$ such that for any $(\mathbf{d}(t))_{t\in\mathbb{N}} \in \mathcal{D}^{\mathbb{N}}$ and all $t \in \mathbb{N}$ we have $||\mathbf{x}(t)|| \geq \kappa$.*

Proof. Choose $\mathbf{x}(0) = \kappa \mathbf{a}_0$, where $\mathbf{a}_0 = \mathbf{a}/||\mathbf{a}||$. Then by using the recursions (7) we get for all $t \in \mathbb{N}$ that

$$\langle \mathbf{a}_0, \mathbf{x}(t) \rangle = \kappa \langle \mathbf{a}_0, \mathbf{a}_0 \rangle - \sum_{i=0}^{t-1} \langle \mathbf{a}_0, \mathbf{d}(t) \rangle \geq \kappa,$$

where the inequality comes from that $\langle \mathbf{a}_0, \mathbf{a}_0 \rangle = 1$ and for all $\mathbf{d}(t) \in \mathcal{D}$ we have

$$\langle \mathbf{a}_0, \mathbf{d} \rangle = (1/||\mathbf{a}||) \langle \mathbf{a}, \mathbf{d} \rangle \leq 0.$$

In particular, the whole sequence $(\mathbf{x}(t))_{t\in\mathbb{N}}$ is in the halfspace $\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^R \big| \langle \mathbf{a}_0, \mathbf{x} \rangle \geq \kappa\}$. Clearly $\mathbf{x}(0) = \kappa \mathbf{a}_0$ is is the optimal solution to

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & ||\mathbf{x}|| \\ \text{subject to} & \mathbf{x} \in \mathcal{H}. \end{array} \tag{13}$$

□

We now provide some examples of θ-covers.

Example 1 (Minimal Example 1: $|\mathcal{D}_1| = R + 1$)**.** *Set*

$$\mathcal{D}_1 = \{\mathbf{e}_1, \cdots, \mathbf{e}_R, -\mathbf{1}/\sqrt{R}\}, \tag{14}$$

where $\mathbf{e}_i$ is the i-th element of the normal basis and $\mathbf{1}$ is R dimensional vector with 1 in every component. Clearly, $|\mathcal{D}| = R + 1$ and therefore $\mathcal{D}$ can be coded using only $\log_2(R + 1)$ bits. We show in Section III-C that this is minimal quantization, since in general if $|\mathcal{D}| \leq R$ then $\mathcal{D}$ can not be a proper quantization.

We show in Lemma 3 in Appendix that $\mathcal{D}_1$ is a θ-cover with

$$\theta = \frac{1}{\sqrt{R^2 + 2\sqrt{R}(R-1)}}. \tag{15}$$

Example 2 (Minimal Example 2: $|\mathcal{D}_2| = R + 1$)**.** *To come later.*

Example 3 ($\pm$ Normal Basis: $|\mathcal{D}_3| = 2R$)**.** *Set*

$$\mathcal{D}_3 = \{\mathbf{e}_1, -\mathbf{e}_1, \mathbf{e}_2, -\mathbf{e}_2, \cdots, \mathbf{e}_R, -\mathbf{e}_R\}.$$

Clearly, $|\mathcal{D}_3| = 2N$ and hence $\log_2(2R)$ bits are needed to broadcast the quantized gradient direction.

Let us now show that $\mathcal{D}_3$ is θ-cover with $\theta = 1/\sqrt{N}$. Take any $\mathbf{x} \in \mathbb{R}^N$ and without loss of generality assume $||\mathbf{x}|| = 1$, since the angle between vectors does not depend of the size of the vectors. Then if we choose

$$\mathbf{d} = sign(\mathbf{x}_i)\mathbf{e}_i \quad \textit{where} \quad i = \textit{argmax}_{i=1,\cdots,R} |\mathbf{x}_i|$$

then it holds that

$$\cos(ang(\mathbf{x}, \mathbf{d})) = \langle \mathbf{x}, \mathbf{d} \rangle = \mathbf{x}_i \cdot sign(\mathbf{x}_i) = |\mathbf{x}_i| \geq \frac{1}{\sqrt{N}}.$$

Example 4 (Signs of the gradients: $|\mathcal{D}_4|{=}2^N$ and $\theta{=}1/\sqrt{N}$)**.** *Set*

$$\mathcal{D} = \{(1/\sqrt{N})(e_1, e_2, \cdots, e_R) \big| e_i \in \{-1, 1\}\}.$$

Here, each $\mathbf{d} \in \mathcal{D}$ represents one orthon of $\mathbb{R}^R$. Therefore, this choose is well suited when the sources can not cooperate and each source updates prize based on local estimate their own part of the gradient, i.e., $\mathbf{d}(t) = sign(\nabla D(\mathbf{p}(t)))$. It is easily checked that $|\mathcal{D}| = 2^N$ hence $\log_2(2^N) = N$ bits are needed to broadcast the quantized gradient direction.

We show that $\mathcal{D}_4$ is θ-cover with $\theta = 1/\sqrt{N}$. Take any $\mathbf{x} \in \mathbb{R}^N$ and without loss of generality assume $||\mathbf{x}|| = 1$. Then if we choose $\mathbf{d} = (1/\sqrt{N}) sign(\mathbf{x})$ then it holds that

$$\begin{aligned} \cos(ang(\mathbf{x}, \mathbf{d})) = \langle \mathbf{x}, \mathbf{d} \rangle &= \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \mathbf{x}_i \cdot sign(\mathbf{x}_i) \\ &\geq \frac{1}{\sqrt{R}} \sum_{i=1}^{R} \mathbf{x}_i^2 = \frac{1}{\sqrt{N}} ||\mathbf{x}|| = \frac{1}{\sqrt{R}}. \end{aligned}$$

Next, we show how Definition 2 can be used to construct a quantized gradient descent algorithm.

B. Answer to Question B)

Unlike Definition 1, Definition 2 acctually provides us with tools to construct algorithms for solving Problem (1) when $\mathcal{D}$ is a proper quantization. In particular, for $\mathbf{p}(t) \in \mathbb{R}$ we can quantize the gradient $\nabla D(\mathbf{p}(t))$ with a $\mathbf{d}(t) \in \mathcal{D}$ such that $\langle \nabla D(\mathbf{p}(t)), \mathbf{d}(t) \rangle \geq \theta$, we demonstrate the step of the algorithm in Algorithm 1. In Section IV, we demonstrate how to choose the step sizes $\gamma(t)$ in Algorithm 1 to achieve different convergence results. In particular, with constant step-size, we show that for any $\epsilon > 0$ we can choose $\gamma(t) = \gamma$ such that $\epsilon > D(\mathbf{p}(t)) - D^\star$ for all t large enough. Moreover, we show that when the step-sizes are non-summable but square summable, i.e.,

$$\sum_{t=0}^{\infty} \gamma(t) = \infty, \quad \sum_{t=0}^{\infty} \gamma(t)^2 < \infty,$$

then $\lim_{t\infty} D(\mathbf{p}(t)) = D^\star$. Note that these results are similar to those obtained for the gradient method (**??**), where the full gradient direction is known.

Algorithm 1: θ-Quantized Gradient Descent (θ-QGD)

Initialization: *Choose* $\mathbf{x}(0) \in \mathbb{R}^N$;
for $t = 0, 1 \cdots$ **do**
  Quantise Gradient: *Choose any* $\mathbf{d}(t) \in \mathcal{D}$ *such that*

$$\langle \nabla f(\mathbf{x}(t)), \mathbf{d}(t) \rangle \geq \theta$$

  Gradient Step: $\mathbf{x}(t+1) = \mathbf{x}(t) - \gamma(t)\mathbf{d}(t)$

C. Asnwer to Question C)

We now demonstrate the connected As noted in, we constructed example of proper quantization sets $\mathcal{D}$ in both Example **??** and **??**.

The intuition behind the result is simple: in $\mathbb{R}^N$ the cone spanned by N vectors is inside a half-space in $\mathbb{R}^N$. We formalize this idea in the following lemma.

Proposition 2. *Suppose that $|\mathcal{D}| \leq N$. Then $\mathcal{D}$ is not a proper quantization (Defintion 1).*

Proof. First consider the case where either $|\mathcal{D}| < N$ or $|\mathcal{D}| = N$ and the elements of $\mathcal{D}$ are linearly dependent. Then $\texttt{Span}(\mathcal{D})$ is a proper subspace of $\mathbb{R}^N$, so there exists a normal $\mathbf{a} \in \mathbb{R}^N$ such that $\langle \mathbf{a}, \mathbf{x} \rangle \leq 0$ for all $\mathbf{x} \in \texttt{Span}(\mathcal{D})$. In particular, $\mathcal{D}$ can not be a theta cover for any $\theta > 0$, since

$$\cos(\texttt{ang}(\mathbf{a}, \mathbf{d})) = \frac{\langle \mathbf{a}, \mathbf{d} \rangle}{\|\mathbf{a}\|\|\mathbf{d}\|} \leq 0.$$

Therefore, the result follows from Proposition 1.

Let us next consider the other case, where $|\mathcal{D}| = N$ and the vectors of $\mathcal{D}$ are linearly independent, i.e., $\texttt{Span}(\mathcal{D}) = \mathbb{R}^N$. Define $\mathbf{D} \in \mathbb{R}^{N \times N}$ such that for $i = 1, \cdots, N$ row i in $\mathbf{D}$ is the i-th elemnt of $\mathcal{D}$, where the elements have some arbitrary order. Then $\mathbf{D}$ is invertable and we can choose $\mathbf{a} = \mathbf{D}^{-1}(-\mathbf{1})$ where $\mathbf{1} \in \mathbb{R}^N$ is a vector of all ones. Then we have for $i = 1, \cdots, N$ that $\langle \mathbf{d}_i, \mathbf{a} \rangle = -\mathbf{d}_i \mathbf{D}^{-1} \mathbf{1} = -1$. Hence, as in the previous case, we get that $\langle \mathbf{a}, \mathbf{x} \rangle \leq 0$ for all $\mathbf{x} \in \texttt{Span}(\mathcal{D})$ implying that $\mathcal{D}$ can not be a theta cover for any $\theta > 0$, and the result follows from Proposition 1. □

D. Answer to Question D)

IV. Convergence

In this section, we provide the convergence results for Algorithm 1. We first investigate the convergence when the step-size is fixed, i.e., $\gamma(t) = \gamma$, in subsection IV-A and then in in subsection IV-B and then provide results for diminishing step-size, i.e., where $\gamma(t) \to 0$ when $t \to \infty$.

A. Constant Step Size

For constant step-size we considering the following two type of stopping criterion

$$\|\nabla D(\mathbf{x})\| < \epsilon, \qquad \text{(Type-1)}$$
$$D(\mathbf{p}(t)) - D^\star < \epsilon. \qquad \text{(Type-2)}$$

We note that when performing primal or dual composition the hence (Type-1) tends to be more practical stopping condition. In particular, for a given ϵ-accuracy, we provide γ and $T \in \mathbb{N}$ that can achieve either Moreover, we provide a tight lower bound on the number of In both cases, we also show that even when the stopping condition is

1) Stopping Condition of (Type-1)*:* We provide couple to capture different engineering needs, e.g., if the best performance for fixed number of iterations, minimal number of iterations for desired accuracy. Let us start

Proposition 3. *For $\epsilon > 0$ we define the set*

$$\mathcal{P}(\epsilon) = \{\mathbf{p} \in \mathbb{R}^R \big| \|\nabla D(\mathbf{p})\| \leq \epsilon\}. \qquad (16)$$

The following holds:

add one result about how theta cooresponsing to the size of quantization

a) For any step size $\gamma \in]0, 2\theta\epsilon/L[$ then there exists $T \in \mathbb{N}$ such that $\mathbf{p}(T) \in \mathcal{P}(\epsilon)$ with T bounded by

$$T \leq \left\lceil \frac{2(D(\mathbf{p}(0)) - D^\star)}{\gamma(2\theta\epsilon - L\gamma)} \right\rceil, \qquad (17)$$

The upper bound (17) *is minimized with the optimal step size $\gamma^\star = \theta\epsilon/L$.*

b) For any step size $\gamma > 0$ and scalar $\kappa > 0$, if we choose $\epsilon = \kappa + \gamma L/(2\theta)$ then there exists $T \in \mathbb{N}$ such that $\mathbf{p}(T) \in \mathcal{P}(\epsilon)$, with T bounded by

$$T \leq \left\lceil \frac{(D(\mathbf{p}(0)) - D^\star)}{\theta\gamma\kappa} \right\rceil, \qquad (18)$$

c) For any tight θ-cover $\mathcal{D}$, the bounds (17) *and* (18) *are tight, i.e., there exists a problem of the form* (1) *where Assumption 1 holds and $\mathbf{p}(0) \in \mathbb{R}^R$ such that $\mathbf{p}(T) \in \mathcal{P}(\epsilon)$ only when the respective bounds* (17) *or* (18) *hold with equality.*

Proof. a) Let $\epsilon > 0$ be given and choose any $\gamma \in]0, 2\theta\epsilon/L[$. From Lemma 4 in the Appendix we get that for all $\mathbf{p}(t) \in \mathbb{R}^R \setminus \mathcal{P}(\epsilon)$ we have

$$D(\mathbf{p}(t+1)) \leq D(\mathbf{p}(t)) - \delta(\epsilon, \gamma, \theta), \qquad (19)$$

where $\delta(\epsilon, \gamma, \theta) > 0$ is defined as in eq. (72). By recursively using (20), it follows that if $\mathbf{p}(t) \in \mathbb{R}^R \setminus \mathcal{P}(\epsilon)$ for $t = 0, \cdots, s-1$ then

$$D(\mathbf{p}(s)) \leq D(\mathbf{p}(0)) - s\ \delta(\epsilon, \gamma, \theta). \qquad (20)$$

Therefore, there must exists $T \leq \lceil (D(\mathbf{p}(0)) - D^\star)/\delta(\epsilon, \gamma, \theta) \rceil$ such that $\mathbf{p}(T) \in \mathcal{P}(\epsilon)$, since otherwise we can use (20) with $s = \lceil (D(\mathbf{p}(0)) - D^\star)/\delta(\epsilon, \gamma, \theta) \rceil + 1$ to get the contradiction that $D(\mathbf{p}(s)) < D^\star$, which can not be true since $D^\star$ is the optimal solution to (1). By rearranging $\lceil (D(\mathbf{p}(0)) - D^\star)/\delta(\epsilon, \gamma, \theta) \rceil$, we get (17).

The optimal step-size $\gamma^\star = \theta\epsilon/L$ comes by simply maximizing the denominator in (17).

We show the result using contradiction and suppose that $\|\nabla D(\mathbf{p}(\mathbf{t}))\| > \epsilon$ for all $T \in \mathbb{N}$. Let us start by using the contraction to provide the following lower bound on $\langle \nabla D(\mathbf{p}(t)), \mathbf{d}(t) \rangle$. For all $t \in \mathbb{N}$ we have

$$\alpha(t) \leq \langle \nabla D(\mathbf{p}(t)), \mathbf{d}(t) \rangle, \qquad (21)$$

where

$$\alpha(t) = \begin{cases} \langle \nabla D(\mathbf{p}(0)), \mathbf{d}(0) \rangle - L\gamma t, & \text{if } t = 1, \cdots, \lfloor T_0 \rfloor, \\ \theta\epsilon, & \text{otherwise,} \end{cases} \qquad (22)$$

and

$$T_0 = \frac{\langle \nabla D(\mathbf{p}(0)), \mathbf{d}(0) \rangle - \theta\epsilon}{\gamma L}. \qquad (23)$$

The bound (21) comes by combining the the following inequalities

$$\langle \nabla D(\mathbf{p}(0)), \mathbf{d}(0) \rangle - L\gamma t \leq \langle \nabla D(\mathbf{p}(t)), \mathbf{d}(t) \rangle, \qquad (24)$$
$$\theta\epsilon \leq \langle \nabla D(\mathbf{p}(t)), \mathbf{d}(t) \rangle, \qquad (25)$$

where (24) is obtained by recursively using

$$\langle \nabla D(\mathbf{x}(t)), \mathbf{d}(t)\rangle \leq L\gamma + \langle \nabla D(\mathbf{p}(t+1)), \mathbf{d}(t)\rangle, \quad (26)$$
$$\leq L\gamma + \langle \nabla D(\mathbf{p}(t+1)), \mathbf{d}(t+1)\rangle, \quad (27)$$

where (26) comes from that D is convex with Lipschitz continuous gradients, see [9, (2.1.9) in Theorem 2.1.5], and (27) comes from that $\mathcal{D}$ is symmetric θ-cover. On the other hand, (25) comes by using that $||\nabla D(\mathbf{p}(t))|| > \epsilon$ for all $t \in \mathbb{N}$. Clearly, $\alpha(t) = \max\{\theta\epsilon, \langle \nabla D(\mathbf{p}(0)), \mathbf{d}(0)\rangle - L\gamma t\}$.

We next use the inequality (21) to get a bound on $D(\mathbf{p}(t)) - D^\star$. By using that the gradients of D are L-Lipschitz continuous, we can apply the descent lemma, see for example [9, eq. (2.1.6)] or [8, Proposition A.24], which states that for all γ we have

$$D(\mathbf{p}(t+1)) \leq D(\mathbf{p}(t)) + \left(\frac{L}{2}\gamma - \langle \nabla D(\mathbf{p}(t), \mathbf{d}(t)\rangle\right)\gamma \quad (28)$$
$$\leq D(\mathbf{p}(t)) + \frac{L}{2}\gamma^2 - \alpha(t)\gamma \quad (29)$$
$$\leq D(\mathbf{p}(0)) + \frac{L}{2}\gamma^2 t - \gamma\sum_{i=0}^{t}\alpha(i) \quad (30)$$
$$\leq D(\mathbf{p}(0)) + \left(\frac{\gamma L}{2} - \theta\epsilon\right)\gamma t \quad (31)$$
$$+ \underbrace{T_0\left(\frac{T_0+1}{2}\gamma L - \langle \nabla D(\mathbf{p}(0)), \mathbf{d}(0)\rangle + \theta\epsilon\right)\gamma}_{=:\Gamma},$$

where (31) comes by noting that

$$\sum_{i=0}^{t}\alpha(i) = (t - \lfloor T_0\rfloor)\theta\epsilon + \sum_{i=0}^{\lfloor T_0\rfloor}(\langle \nabla D(\mathbf{p}(0), \mathbf{d}(0)\rangle - L\gamma i)$$
$$= t\theta\epsilon - \lfloor T_0\rfloor\left(\frac{\lfloor T_0\rfloor + 1}{2}\gamma L + \theta\epsilon - \langle \nabla D(\mathbf{p}(0), \mathbf{d}(0)\rangle\right).$$

By substituting $D^\star$ on both sides of (31) we obtain

$$D(\mathbf{p}(t+1)) - D^\star \leq D(\mathbf{p}(0)) - D^\star + \Gamma + \left(\frac{\gamma L}{2} - \theta\epsilon\right)\gamma t.$$

In particular, it follows that

$$D(\mathbf{p}(t+1)) - D^\star < 0$$

for all

$$t \geq \frac{2(D(\mathbf{p}(0)) - D^\star)}{(2\theta\epsilon - \gamma L)} \quad (32)$$

recalling that $D(\mathbf{p}(t)) - D^\star \geq 0$

b) This results can be obtained by using almost identical arguments as used to prove part a). The only difference is that now we have explicit form for ϵ when using (20), which results in

$$\delta(\epsilon, \gamma, \theta) = \delta\left(\kappa + \frac{\gamma L}{2\theta}, \gamma, \theta\right) = \theta\gamma\kappa. \quad (33)$$

□

Another interesting is where the This motivates the following result, where we provide the optimal guaranteed accuracy for provided bound $T^{\max}$ on the number of iterations. B

Proposition 4. *Let an upper bound $T^{\max} \in \mathbb{N}$ on number of iterations be given. Then the optimal accuracy $\epsilon^\star$ that can be achivieved and the associated step size $\gamma^\star > 0$ are given by*

$$\epsilon^\star = \frac{L}{\theta}\gamma^\star \quad \text{and} \quad \gamma^\star = \sqrt{\frac{2(D(\mathbf{p}(0)) - D^\star)}{LT}}. \quad (34)$$

In other words, if we chose there is and for any $\epsilon < \epsilon^\star$ there exists a problem such that $\mathbf{p}(t) \notin \mathcal{P}(\epsilon)$ for $t = 0, \cdots, T^{\max}$ and all $\gamma > 0$.

Note: The following result includes some of the same results as above, but also tells something about the Type-II stopping condition.

Proposition 5. *Set $\mathcal{X}(\epsilon) = \{\mathbf{x} \in \mathbb{R}^N | ||\nabla f(\mathbf{x})|| \leq \epsilon\}$. For any $\epsilon > 0$, if $\gamma \in]0, \frac{2\epsilon}{L\sqrt{N}}[$ then the following holds.*

a) For all $\mathbf{x}(t) \in \mathbb{R}^N \setminus \mathcal{X}(\epsilon)$ we get

$$f(\mathbf{x}(t+1)) \leq f(\mathbf{x}(t)) - \delta, \quad (35)$$

where $\delta = -\gamma\left(\frac{L}{2}\gamma - \frac{\epsilon}{\sqrt{N}}\right) > 0$.

b) There exists $T \in \mathbb{N}$ such that $\mathbf{x}(T) \in \mathcal{X}(\epsilon)$ and

$$T \leq \left\lceil \frac{f(\mathbf{x}(0)) - f^\star}{\delta} \right\rceil,$$

where $f^\star$ is the optimal value to (1).

c) For the T in b) it holds all $t \geq T$ that

$$f(\mathbf{x}(t)) \leq f^\epsilon + \left(\epsilon + \frac{L}{2}\gamma\right)\gamma$$

for all $t \geq T$, where

$$f^\epsilon = \max\{f(\mathbf{x}) | \mathbf{x} \in \mathcal{X}(\epsilon)\} \quad (36)$$

and $\lim_{t\to\infty} f^\epsilon = f^\star$.

d) If f is also μ-strongly convex and T is chosen as in b) then for all $t \geq T$ we have

$$f(\mathbf{x}(t)) \leq f^\star + \frac{\epsilon^2}{2\mu} + \left(\epsilon + \frac{L}{2}\gamma\right)\gamma. \quad (37)$$

Proof. a) Take arbitrary $\epsilon > 0$ and $\mathbf{x}(t) \in \mathbb{R}^N \setminus \mathcal{X}(\epsilon)$. Then by using that the gradients are L-Lipschitz continuous, we can apply the descent lemma, see for example [9, eq. (2.1.6)] or [8, Proposition A.24], which states that for all γ we have

$$f(\mathbf{x}(t) - \gamma\mathbf{d}(t)) \leq f(\mathbf{x}(t)) - \langle \nabla f(\mathbf{x}(t), \mathbf{d}(t)\rangle\gamma + \frac{L}{2}||\mathbf{d}(t)||^2\gamma^2,$$
$$= f(\mathbf{x}(t)) - \langle \nabla f(\mathbf{x}(t)), \mathbf{d}(t)\rangle\gamma + \frac{L}{2}\gamma^2,$$
$$= f(\mathbf{x}(t)) + \left(\frac{L}{2}\gamma - \langle \nabla f(\mathbf{x}(t)), \mathbf{d}(t)\rangle\right)\gamma, \quad (38)$$

where the equality comes by using that $||\mathbf{d}(t)|| = 1$. We also have

$$\frac{L}{2}\gamma - \langle \nabla f(\mathbf{x}(t), \mathbf{d}(t)\rangle \leq \frac{L}{2}\gamma - \frac{\epsilon}{\sqrt{N}} \quad (39)$$
$$< 0, \quad (40)$$

where (39) comes by using Lemma **??** together with that $\mathbf{x}(t) \in \mathbb{R}^N \setminus \mathcal{X}(\epsilon)$ to get

$$\begin{aligned}\langle \nabla f(\mathbf{x}(t), \mathbf{d}(t)\rangle =& ||\nabla f(\mathbf{x}(t))|| \cos(\texttt{ang}(\mathbf{d}(t), \nabla f(\mathbf{x}(t)))) \\ \geq& \frac{\epsilon}{\sqrt{N}}, \end{aligned} \tag{41}$$

and (40) comes by that the step-size is choosen such that $\gamma < \frac{2\epsilon}{L\sqrt{N}}$. Using (40) we see that

$$\delta = -\gamma\left(\frac{L}{2}\gamma - \frac{\epsilon}{\sqrt{N}}\right) > 0, \tag{42}$$

and by combining (74), (39), and (40) get (35).

b) We show this by contradiction. Suppose that $\mathbf{x}(t) \in \mathbb{R}^N \setminus \mathcal{X}(\epsilon)$ for all $t \leq T_0$, where

$$T_0 = \left\lceil \frac{f(\mathbf{x}(0)) - f^\star}{\delta} \right\rceil.$$

Then from Part a) we have that

$$f(\mathbf{x}(t)) - f^\star \leq (f(\mathbf{x}(0)) - f^\star) - t\delta,$$

for all $t \geq 0$. In particular at time T_0, we have that $f(\mathbf{x}(T_0)) - f^\star \leq 0$ which implies that $f(\mathbf{x}(T_0)) = f^\star \leq 0$, and $\mathbf{x}(T_0)$ is an optimal solution of (1). Therefore, by the optimality conditions, we must have $\nabla f(\mathbf{x}(T_0)) = \mathbf{0}$, which contradicts that $\mathbf{x}(T_0) \in \mathbb{R}^N \setminus \mathcal{X}(\epsilon)$.

c) Choose $T \in \mathbb{N}$ such that $\mathbf{x}(T) \in \mathcal{X}(\epsilon)$, such a T exist from b). Then for all $t \geq T$ such that $\mathbf{x}(t) \in \mathcal{X}(\epsilon)$ we have from [9, eq. (2.1.7)] that

$$\begin{aligned} f(\mathbf{x}(t+1)) \leq& f(\mathbf{x}(t)) + \langle \nabla f(\mathbf{x}(t)), -\gamma \mathbf{d}(t)\rangle + \frac{L}{2}\gamma^2 && (43)\\ =& f(\mathbf{x}(t)) + \epsilon\gamma + \frac{L}{2}\gamma^2 && (44)\\ =& f(\mathbf{x}(t)) + \left(\epsilon + \frac{L}{2}\gamma\right)\gamma && (45)\\ \leq& f^\epsilon + \left(\epsilon + \frac{L}{2}\gamma\right)\gamma && (46)\end{aligned}$$

where the equality comes by using that $||\nabla f(\mathbf{x}(t))|| \leq \epsilon$ and that $||\mathbf{d}(t)|| = 1$. Then, for all $\mathbf{x}(t) \in \mathbb{R}^N \setminus \mathcal{X}(\epsilon)$ we have from a) that $f(\mathbf{x}(t+1)) < f(\mathbf{x}(t))$. Hence, we must have that $\mathbf{x}(t) \leq f^\epsilon + \left(\epsilon + \frac{L}{2}\gamma\right)\gamma$ for all $t \geq T$.

d) For any $\mathbf{x}(t) \in \mathcal{X}(\epsilon)$ we have that

$$\begin{aligned} f(\mathbf{x}(t)) \leq& f^\star + \frac{1}{2\mu}||\nabla f(\mathbf{x}(t))||^2, && (47)\\ \leq& f^\star + \frac{\epsilon^2}{2\mu}, && (48)\end{aligned}$$

where (47) comes from using that f is μ-strongly convex and [9, eq. (2.1.19) in Theorem 2.1.10] and that $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ for all $\mathbf{x}^\star \in \mathcal{P}^\star$. □

B. Diminishing Step Size

Proposition 6. *Suppose that* $\lim_{t\to\infty}\gamma(t) = 0$*,*

$$\sum_{t=0}^{N}\gamma(t) = \infty, \quad \textit{and} \ \sum_{t=0}^{N}\gamma(t)^2 < \infty, \tag{49}$$

then

$$\lim_{t\to\infty} f(\mathbf{x}(t)) = f^\star. \tag{50}$$

Proof. Define $\mathcal{X}(\epsilon)$ as in Proposition 5. Then, from Proposition 5 a), for any $\epsilon > 0$ we can find $\bar{\gamma} > 0$ such that for all $\gamma \in]0, \bar{\gamma}]$ and $\mathbf{x}(t) \in \mathbb{R}^N \setminus \mathcal{X}(\epsilon)$ we have that $\gamma\left(\frac{L}{2}\gamma - \frac{\epsilon}{\sqrt{N}}\right) < 0$ and

$$f(\mathbf{x}(t) - \gamma\mathbf{d}(t)) \leq f(\mathbf{x}(t)) + \gamma\left(\frac{L}{2}\gamma - \frac{\epsilon}{\sqrt{N}}\right). \tag{51}$$

Let us now show by contradiction that for all $\epsilon > 0$ there exist infinity many $T \in \mathbb{N}$ such that $\mathbf{x}(T) \in \mathcal{X}(\epsilon)$. Suppose the contrary, that there is $\epsilon > 0$ and $T_0 \in \mathbb{N}$ such that $\mathbf{x}(t) \in \mathbb{R}^N \setminus \mathcal{X}(\epsilon)$ for all $t \geq T_0$. Then since $\lim_{t\to\infty}\gamma(t) = 0$, we can choose $T \in \mathbb{N}$ such that $T \geq T_0$ and $\gamma(t) \in]0, \bar{\gamma}]$ for all $t \geq T$, and hence by (51) we have for all $t \geq T$ that $\gamma(t)\left(\frac{L}{2}\gamma(t) - \frac{\epsilon}{\sqrt{N}}\right) < 0$ and

$$\begin{aligned} f(\mathbf{x}(t+1)) \leq& f(\mathbf{x}(t)) + \gamma(t)\left(\frac{L}{2}\gamma(t) - \frac{\epsilon}{\sqrt{N}}\right), && (52)\\ \leq& f(\mathbf{x}(T)) + \frac{L}{2}\sum_{\tau=T}^{t}\gamma(\tau)^2 - \frac{\epsilon}{\sqrt{N}}\sum_{\tau=T}^{t}\gamma(\tau). && (53)\end{aligned}$$

Now since $\gamma(t)$ is non-summable but square-summable (cf. (49)), the right hand side of (53) diverges to $-\infty$ implying that $\lim_{t\to\infty} f(\mathbf{x}(t)) = -\infty$. However, $\liminf_{t\to\infty} f(\mathbf{x}(t)) > -\infty$ $f(\mathbf{x}(t)) \leq f^\star > -\infty$ by Assumption 1, contradicting $\lim_{t\to\infty} f(\mathbf{x}(t)) = -\infty$.

Let us now show that for any $\epsilon > 0$ there exists $T \in \mathbb{N}$ such that $f(\mathbf{x}(t)) - f^\star < \epsilon$ for all $t \geq T$, i.e., $\lim_{t\to\infty} f(\mathbf{x}(t)) = f^\star$. Choose $\kappa = \min\{\kappa_1, \kappa_2\}$ where

$$\begin{aligned} \kappa_1 =& \frac{1}{4}\sqrt{\frac{\epsilon LN}{1+\sqrt{N}}}, && (54)\\ f^{\kappa_2} <& f^\star + \frac{\epsilon}{2}, && (55)\end{aligned}$$

where

$$f^{\kappa_2} = \max\{f(\mathbf{x}) \big| ||f(\mathbf{x})|| \leq \kappa_2\}, \tag{56}$$

such a κ_2 exists since ∇f is continues and $f(\mathbf{x}) = f^\star$ for all $\mathbf{x}$ such that $\nabla f(\mathbf{x}) = \mathbf{0}$. Now, from above, there exist infinity many $T \in \mathbb{N}$ such that $\mathbf{x}(T) \in \mathcal{X}(\kappa)$, hence since $\lim_{t\to\infty}\gamma(t) = 0$ we can choose $T \in \mathbb{N}$ such that $\mathbf{x}(T) \in \mathcal{X}(\kappa)$ and $\gamma(t) < \frac{2\kappa}{L\sqrt{N}}$ for all $t \geq T$. Hence, from Proposition 5 c) for all $t \geq T$ we have that

$$\begin{aligned} f(\mathbf{x}(t)) \leq& f^\kappa + \left(\kappa + \frac{\kappa}{\sqrt{N}}\right)\frac{2\kappa}{L\sqrt{N}} && (57)\\ =& f^\kappa + \frac{2\kappa^2(1+\sqrt{N})}{LN} && (58)\\ \leq& f^\star + \frac{\epsilon}{2} + \frac{\epsilon}{2} = f^\star + \epsilon, && (59)\end{aligned}$$

where (59) comes by using (54) and (55).

□

Proposition 7. *Define* $T = \min\{t \in \mathbb{N} | \mathbf{x}(t) \in \mathcal{X}(\epsilon)\}$ *then for* $t = 0, \cdots T$*, want to say something about convergence rate. The main problem is that even though we can show*

Proof. For all $\mathbf{x}(t) \in \mathcal{X}(\epsilon)$ we have

$$f(\mathbf{x}(t+1)) \leq f(\mathbf{x}(t)) + \gamma \left(\frac{\gamma L}{2} - \frac{||\nabla f(\mathbf{x}(t))||}{\sqrt{N}} \right), \tag{60}$$

$$\leq f(\mathbf{x}(t)) + \underbrace{\gamma \left(\frac{\gamma L}{2\epsilon} - \frac{1}{\sqrt{N}} \right)}_{=:\omega} ||\nabla f(\mathbf{x}(t))||, \tag{61}$$

where (60) comes by using (74) and Lemma **??** and (61) comes from that $\mathbf{x}(t) \in \mathcal{X}(\epsilon)$. We also have that

$$a \tag{62}$$

□

APPENDIX

Lemma 3. *Recall $\mathcal{D}_1$ defined in eq.* (14)*, Example 1, Section III-A. $\mathcal{D}_1$ is a θ-cover with the θ defined in eq.* (15)*.*

Proof. To obtain the desired result, we show that for θ in (15) it holds that for any $\mathbf{x} \in \mathcal{S}^R$ there exists $\mathbf{d} \in \mathcal{D}_1$ such that (2) holds.

First consider the case where $\mathbf{x}_j \geq \theta$ for some component j. Then if we choose $\mathbf{e}_j \in \mathcal{D}_1$ we get

$$\cos(\texttt{ang}(\mathbf{x}, \mathbf{e}_j)) = \langle \mathbf{x}, \mathbf{e}_j \rangle = \mathbf{x}_j \geq \theta.$$

Therefore, we finalize the proof by showing that if $\mathbf{x} \in \mathcal{S}^R$ and $\mathbf{x}_i \leq \theta$ for $i = 1, \cdots, R$ then

$$\cos\left(\texttt{ang}\left(\mathbf{x}, -\frac{1}{\sqrt{R}}\mathbf{1} \right) \right) = -\frac{1}{\sqrt{R}} \sum_{i=1}^{R} \mathbf{x}_i \geq \theta,$$

Without loss of generality, let the components of $\mathbf{x}$ be ordered so that

$$\begin{aligned} \mathbf{x}_i &\geq 0, \quad \text{if } i = 1, \cdots, K, \text{ and} \\ \mathbf{x}_i &< 0, \quad \text{if } i = K+1, \cdots, R \end{aligned}$$

where K is the number of positive components of $\mathbf{x}$. Then it holds that

$$-\frac{1}{\sqrt{R}} \sum_{i=1}^{R} \mathbf{x}_i \geq -\frac{1}{\sqrt{R}} \left(\sum_{i=1}^{K} \mathbf{x}_i - \sqrt{1 - \sum_{i=1}^{K} \mathbf{x}_i^2} \right) \tag{63}$$

$$\geq -\frac{1}{\sqrt{R}} \left(\sum_{i=1}^{K} \theta - \sqrt{1 - \sum_{i=1}^{K} \theta^2} \right) \tag{64}$$

$$= -\frac{1}{\sqrt{R}} \left(K\theta - \sqrt{1 - K\theta^2} \right), \tag{65}$$

where (63) comes by using that $\sum_{i=1}^{R} \mathbf{x}_i^2 = 1$ and the inequality between the 1 and 2 norm, i.e.,

$$\sum_{i=K+1}^{R} |\mathbf{x}_i| \geq \sqrt{\sum_{i=K+1}^{R} \mathbf{x}_i^2} = \sqrt{1 - \sum_{i=1}^{K} \mathbf{x}_i^2},$$

and that $\mathbf{x}_i < 0$, for $i > K$, and (64) comes by noting that (63) is decreasing and that $\mathbf{x}_i \leq \theta$ for all i. Now, by inserting our choose of θ, see (15), in (65) we get

$$-\frac{1}{\sqrt{R}} \sum_{i=1}^{R} \mathbf{x}_i \geq -\frac{1}{\sqrt{R}} \left(\frac{K}{\sqrt{R^2 + 2\sqrt{R}(R-1)}} - \sqrt{\frac{R^2 + 2\sqrt{R}(R-1) - K}{R^2 + 2\sqrt{R}(R-1)}} \right) \tag{66}$$

$$= -\frac{K - \sqrt{R^2 + 2\sqrt{R}(R-1) - K}}{\sqrt{R}\sqrt{R^2 + 2\sqrt{R}(R-1)}} \tag{67}$$

$$\geq -\frac{(R-1) - \sqrt{R^2 + 2\sqrt{R}(R-1) - (R-1)}}{\sqrt{R}\sqrt{R^2 + 2\sqrt{R}(R-1)}} \tag{68}$$

$$= \frac{\sqrt{R}}{\sqrt{R}\sqrt{R^2 + 2\sqrt{R}(R-1)}} \tag{69}$$

$$= \theta, \tag{70}$$

where (68) comes from that (67) is decreasing in K and $K \leq R - 1$, and (69) comes by using that

$$R^2 + 2\sqrt{R}(R-1) - (R-1) = ((R-1) + \sqrt{R})^2.$$

□

s

Lemma 4. *Let $D : \mathbb{R}^R \to \mathbb{R}$ be a convex and continuously differentiable function with L-continuous gradient. Suppose $\epsilon > 0$, $\gamma \in]0, 2\theta\epsilon/L[$, and $\theta \in]0, 1]$ and $\mathbf{p} \in \mathbb{R}^R \setminus \mathcal{P}(\epsilon)$ and $\mathbf{d} \in \mathcal{S}^R$ where $\cos(ang(\mathbf{x}, \mathbf{d})) \geq \theta$. Then it holds that*

$$D(\mathbf{p} - \gamma \mathbf{d}) \leq D(\mathbf{p}) - \delta(\epsilon, \gamma, \theta) \tag{71}$$

where

$$\delta(\epsilon, \gamma, \theta) = -\left(\frac{L}{2}\gamma - \theta\epsilon \right) \gamma > 0. \tag{72}$$

Proof. By using that the gradients of D are L-Lipschitz continuous, we can apply the descent lemma, see for example [9, eq. (2.1.6)] or [8, Proposition A.24], which states that for all γ we have

$$D(\mathbf{p} - \gamma \mathbf{d}) \leq D(\mathbf{p}) - \langle \nabla D(\mathbf{p}), \mathbf{d} \rangle \gamma + \frac{L}{2} ||\mathbf{d}||^2 \gamma^2, \tag{73}$$

$$= D(\mathbf{p}) + \left(\frac{L}{2}\gamma - \langle \nabla D(\mathbf{p}), \mathbf{d}(t) \rangle \right) \gamma, \tag{74}$$

$$\leq D(\mathbf{p}) + \left(\frac{L}{2}\gamma - \theta\epsilon \right) \gamma \tag{75}$$

$$= D(\mathbf{p}) - \delta(\epsilon, \gamma, \theta) \tag{76}$$

where (74) comes from using that $||\mathbf{d}|| = 1$, (75) comes from using $\cos(\texttt{ang}(\mathbf{x}, \mathbf{d})) \geq \theta$, $||\nabla D(\mathbf{p})|| \geq \epsilon$ (since $\mathbf{p} \in \mathbb{R}^R \setminus \mathcal{P}(\epsilon)$, and that

$$\langle \nabla D(\mathbf{p}), \mathbf{d} \rangle = ||\nabla D(\mathbf{p})|| \cos(\texttt{ang}(\mathbf{d}, \nabla D(\mathbf{p}))).$$

The fact that $\delta(\epsilon, \gamma, \theta) > 0$ is a direct consequence of the choose $\gamma \in]0, 2\theta\epsilon/L[$. □

REFERENCES

[1] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *The Journal of the Operational Research Society*, vol. 49, no. 3, pp. pp. 237–252, 1998. [Online]. Available: http://www.jstor.org/stable/3010473

[2] S. Low and D. Lapsley, "Optimization flow control. i. basic algorithm and convergence," *Networking, IEEE/ACM Transactions on*, vol. 7, no. 6, pp. 861–874, Dec 1999.

[3] D. Lapsley and S. Low, "Random early marking for internet congestion control," in *Global Telecommunications Conference, 1999. GLOBECOM '99*, vol. 3, 1999, pp. 1747–1752 vol.3.

[4] J. Herdtner and E. Chong, "Analysis of a class of distributed asynchronous power control algorithms for cellular wireless systems," *Selected Areas in Communications, IEEE Journal on*, vol. 18, no. 3, pp. 436–446, March 2000.

[5] M. Rabbat and R. Nowak, "Quantized incremental algorithms for distributed optimization," *Selected Areas in Communications, IEEE Journal on*, vol. 23, no. 4, pp. 798–808, April 2005.

[6] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 8, pp. 1439–1451, Aug 2006.

[7] ——, "Alternative distributed algorithms for network utility maximization: Framework and applications," *Automatic Control, IEEE Transactions on*, vol. 52, no. 12, pp. 2254–2269, Dec 2007.

[8] D. P. Bertsekas, *Nonlinear Programming: 2nd Edition*. Athena Scientific, 1999.

[9] Y. Nesterov, *Introductory Lectures on Convex Optimization*. Springer, 2004.