DISTRIBUTED OPTIMIZATION IN AN ENERGY-CONSTRAINED NETWORK

Alireza Razavi and Zhi-Quan Luo

Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455, U.S.A.
Email: {raza0007, luozq}@umn.edu

ABSTRACT

We consider a distributed optimization problem whereby two nodes S_1, S_2 wish to jointly minimize a common convex quadratic cost function $f(x_1, x_2)$, subject to separate local constraints on x_1 and x_2, respectively. Suppose that node S_1 has control of variable x_1 only and node S_2 has control of variable x_2 only. The two nodes locally update their respective variables and periodically exchange their values over a noisy channel. Previous studies of this problem have mainly focused on the convergence issue and the analysis of convergence rate. In this work, we focus on the communication energy and study its impact on convergence. In particular, we consider a class of distributed stochastic gradient type algorithms implemented using certain linear analog messaging schemes. We study the minimum amount of communication energy required for the two nodes to compute an ϵ-minimizer of $f(x_1, x_2)$ in the mean square sense. Our analysis shows that the communication energy must grow at least at the rate of $\Omega(\epsilon^{-1})$. We also derive specific designs which attain this minimum energy bound, and provide simulation results that confirm our theoretical analysis. Extension to the multiple node case is described.

***Index Terms*—** Distributed optimization, Sensor networks, Energy constraint, Stochastic gradient method, Convergence

1. INTRODUCTION

Consider a network of n nodes which collaborate to minimize a cost function $f(x_1, x_2, \ldots, x_n)$, subject to separate constraints on x_i, where x_i is a local (vector) variable controlled by node S_i. Each node can perform local computation and exchange messages with a set of predefined neighbors through orthogonal noisy channels. Moreover, we assume $f(x_1, x_2, \ldots, x_n)$ has a certain *"local structure"* in the sense that its partial derivative with the respect to x_i only depends on the local variables at node S_i and its neighbors.

A distributed optimization problem of this kind arises naturally in sensor network applications. For example, in the sensor localization problem, we are given the locations of anchor nodes and distance measurements between certain neighbor nodes in the network. The goal is to estimate the locations of all sensors in the network by distributed minimization of a cost function $f(x_1, x_2, \ldots, x_n)$ defined by the L_1 norm of distance errors [2]. In this context, x_i is the location of sensor S_i and is to be estimated by S_i. The location x_i may be required to satisfy some local constraints representing a priori information (e.g. range) available at sensor S_i. To minimize $f(x_1, x_2, \ldots, x_n)$, sensor S_i periodically updates its local variable, x_i, and exchanges information with neighbor nodes through orthogonal noisy channels. A special feature of this problem is the fact that nodes are battery operated and hence energy-constrained. Note that energy of each node is consumed for various operations including local computation and inter-sensor communication, with the latter being the dominant part. This motivates us to study the minimum amount of communication energy required for distributed optimization.

Energy consumption has not been a consideration of algorithm design in classical distributed optimization [1]. Even recent studies of distributed optimization in the context of sensor networks [6, 3] have mainly focused on convergence issues such as convergence criteria and convergence rate. To the best of our knowledge, the most relevant work to this paper is [5] which studied the minimum number of bits that must be exchanged between two nodes in order to find an ϵ-minimizer of f. However, unlike our current work the communication channel is assumed distortion-less in [5], and there was no effort to characterize minimum energy consumption.

The main contributions of this paper are twofold. First, we establish an asymptotic lower bound for communication energy required to obtain an ϵ-minimizer of f. Second, we provide specific designs which attain this minimum energy bound. We start with a two node case which is later generalized to the multiple node case. The considered cost function is quadratic and convex. The generalization of this work to a general cost function (e.g $f(x_1, x_2, \ldots, x_n)$ in the stated sensor localization problem) is a subject of ongoing research.

We adopt the following notations: all matrices and vectors will be denoted by upper case characters and bold lower case, respectively. For any variable x_i, we use $x_i^j(t)$ to indicate the value of x_i at node S_j and iteration t. An asymptotic lower bound and asymptotically tight bound will be denoted by Ω and Θ, respectively.

2. ALGORITHM FRAMEWORK

We consider a distributed optimization problem with n nodes, $S_i, i \in \{1, \ldots, n\}$, jointly minimizing a convex quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + \mathbf{c}$, $\mathbf{x} = [x_1, \ldots, x_n]^T$, where $A \in R^{n\times n}$, $A \succ 0$ and $\mathbf{b}, \mathbf{c} \in R^{n\times 1}$. Node S_i has control of scalar variable x_i and knows A_i and b_i which are the i^{th} row of A and $\mathbf{b}$ for $i \in \{1, \ldots, n\}$, respectively. Nodes communicate through orthogonal time-invariant noisy channels using analog messaging. The communication channel between nodes S_i and S_l is corrupted by additive noise, $n^{i,l}(t)$, with zero mean and variance $\sigma_{i,l}^2$. In this model, the received signal by node S_l from S_i, $r^{i,l}(t)$, is

$$r^{i,l}(t) = y^i(t) + n^{i,l}(t), \tag{1}$$

This work is supported by the USDOD Army, grant number W911NF-05-1-0567.

where $y^i(t)$ is the signal transmitted by node S_i. Therefore, communication energy at iteration t is proportional to $E[\|y^i(t)\|^2]$ where E is the expectation operator. This paper aims to derive the communication energy required to obtain an ϵ-minimizer of $f(\mathbf{x})$ in the mean square sense. A point $\mathbf{x}$ is an ϵ-minimizer of $f(\mathbf{x})$ in the mean square sense if $E[\|\mathbf{x}-\mathbf{x}^*\|^2] \leq \epsilon$, where $\mathbf{x}^*$ is the optimum point. In the sequel, we start with a two node case and will extend the results to the multiple node case in Section 3.1.

A distributed algorithm consists of two parts: a communication scheme and a local computation scheme at each node.

A. Communication scheme: After each local update, node $S_i, i \in \{1,2\}$ should relay its information to the other node. One way is to directly send the updated value of its local variable, $x_i^i(t+1)$, resulting in a communication power of $E[\|x_i^i(t+1)\|^2]$ which converges to a constant value. An alternative way is to send the incremental value, $x_i^i(t+1) - x_i^i(t)$, in which case the communication power, $E[\|x_i^i(t+1) - x_i^i(t)\|^2]$, would vanish if x_i^i converges. In general, we can consider a linear analog messaging scheme where the transmitted signal, $y^i(t)$, is given as,

$$y^i(t) = [x_i^i(t) - \gamma(t) x_i^i(t-1)], t = 1, \ldots. \tag{2}$$

In this equation, $\gamma(t)$ is a positive coefficient to be chosen ($\gamma(1) = 0$). Therefore, the total communication energy of S_i is

$$E^i_{com}(T) = \sum_{t=1}^{T} E\|x_i^i(t) - \gamma(t) x_i^i(t-1)\|^2. \tag{3}$$

Equation (3) shows that in order to reduce communication energy, $\gamma(t)$ should converge to 1. Note that the choice of $\gamma(t)$ must also ensure the convergence of the distributed algorithm.

B. Local computation scheme: Optimization algorithms in the presence of noise can be performed based on the stochastic gradient type algorithm or the so-called Robbins-Monro algorithm [4]. One iteration of this algorithm is given as,

$$\mathbf{x}(t+1) = \mathbf{x}(t) - a(t) g(\mathbf{x}),\ t = 1, \ldots,$$

where $a(t)$ is a vanishing positive step size satisfying $\sum_{t=1}^{\infty} a(t) \to \infty; \sum_{t=1}^{\infty} (a(t))^2 < \infty$, and $g(\mathbf{x})$ is a noisy version of the gradient vector of $f(\mathbf{x})$. We consider a distributed implementation of the stochastic gradient type algorithm whereby $S_i, i \in \{1,2\}$ tracks the other node variable according to

$$x_j^i(t+1) = \gamma(t) x_j^i(t) + r^{j,i}(t). \tag{4}$$

Here $r^{j,i}(t)$ is the received message from node S_j at iteration t as defined by (1)-(2). Node S_i uses this noisy copy to estimate $g_i(\mathbf{x})$, the partial derivative of $f(\mathbf{x})$ with the respect of its local variable x_i^i, and update x_i^i as

$$\begin{aligned} &\text{at node } S_1: \\ &x_1^1(t+1) = x_1^1(t) - \tfrac{1}{t+1} g_1\left(x_1^1(t), x_2^1(t+1)\right) \\ &\text{at node } S_2: \\ &x_2^2(t+1) = x_2^2(t) - \tfrac{1}{t+1} g_2\left(x_1^2(t+1), x_2^2(t)\right). \end{aligned} \tag{5}$$

Here, we have chosen $a(t) = \frac{1}{t+1}$. In the next section, we derive a convergence condition on $\gamma(t)$ which will be used to bound the total communication energy.

3. A CONVERGENCE CONDITION AND REQUIRED COMMUNICATION ENERGY

Define $\alpha(j,t) := \sum_{i=j}^{t} \frac{\Theta\left(\frac{i+1}{t+1}\right)^{\lambda_{min}}}{i+1} \prod_{k=j+1}^{i} \gamma(k)$, where λ_{min} is the smallest eigenvalue of the matrix A, and $\prod_{k=i}^{j} \gamma(k) := 1$ for $i > j$.

Theorem 1

(a) *The distributed algorithm described by* (2), (4), (5) *converges to the global minimum of $f(\mathbf{x})$ in the mean square sense when $\gamma(k)$ satisfies*

$$\lim_{t\to\infty} \sum_{i=0}^{t} (\alpha(i,t))^2 \to 0. \tag{6}$$

(b) *To obtain an ϵ-minimizer of $f(\mathbf{x})$ in the mean square sense, the required communication energy must be at least $\Omega(\epsilon^{-1})$.*

Proof (a) We start with the iterative update in equation (5) and obtain a tight bound for the mean square error in terms of the initial values of local variables and channel noise variance. Then, we show that this bound converges to zero under condition (6). We will need the following lemma which is stated here without a proof,

Lemma 1 *For any positive, decreasing integrable function $h(x)$, there holds*

$$\int_i^{t+1} h(x)dx \leq \sum_{j=i}^{t} h(j) \leq h(i) + \int_i^t h(x)dx \leq \int_{i-1}^{t} h(x)dx.$$

Define error at the $(t+1)^{th}$ iteration as $\mathbf{e}(t+1) := \mathbf{x}(t+1) - \mathbf{x}^*$, where $\mathbf{x}(t+1) := [x_1^1(t+1), x_2^2(t+1)]$ is the vector of local variables in respective nodes at $(t+1)^{th}$ iteration. Using equations (1)-(2), (4)-(5), we can show

$$\mathbf{e}(t+1) = \left(I - \frac{A}{t+1}\right)\mathbf{e}(t) + \frac{1}{t+1}\mathbf{v}(t).$$

In this equation, the accumulated channel noise, $\mathbf{v}(t)$, is given as,

$$\mathbf{v}(t) = -\begin{bmatrix} a_{1,2} \sum_{i=1}^{t} \prod_{k=i+1}^{t} \gamma(k) n^{2,1}(i) \\ a_{1,2} \sum_{i=1}^{t} \prod_{k=i+1}^{t} \gamma(k) n^{1,2}(i) \end{bmatrix},$$

where $a_{i,j}$ is the $(i,j)^{th}$ entry of matrix A, and we used the fact that matrix A is symmetric. Consider the eigenvalue decomposition of matrix A:

$$\Lambda := \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = P^T A P,$$

where $\lambda_i, i \in \{1,2\}$ are the eigenvalues of matrix A. For $\bar{\mathbf{e}}(t) := P^T \mathbf{e}(t)$, and $\bar{\mathbf{v}}(t) := P^T \mathbf{v}(t)$, the error at the $(t+1)^{th}$ iteration is

$$\mathbf{e}(t+1) = P\prod_{j=1}^{t}(I - \frac{\Lambda}{j+1})\bar{\mathbf{e}}(1) - \sum_{i=1}^{t} \frac{P\prod_{j=i+1}^{t}\left(I - \frac{\Lambda}{j+1}\right)\bar{\mathbf{v}}(i)}{i+1}. \tag{7}$$

For $h(x) = -\log(1 - \frac{\lambda_l}{x})$ and $2\lambda_l \leq i \leq t$, it follows from Lemma 1 that:

$$\prod_{j=i}^{t}\left(1 - \frac{\lambda_l}{j+1}\right) = \Theta\left(\frac{i}{t+1}\right)^{\lambda_l} \tag{8}$$

The mean square error at the $(t+1)^{th}$ iteration can be written as

$$\begin{aligned} E[\|\mathbf{e}(t+1)\|^2] &\overset{(a)}{=} \Theta\left(\tfrac{1}{t+1}\right)^{2\lambda_{min}} + E\left[\left(\sum_{l=1}^{2}\sum_{i=1}^{t} \frac{\Theta\left(\frac{i+1}{t+1}\right)^{\lambda_{min}} v_l(i)}{i+1}\right)^2\right] \\ &\overset{(b)}{=} \Theta\left(\tfrac{1}{t+1}\right)^{2\lambda_{min}} + 2(a_{1,2})^2\sigma^2 \sum_{j=1}^{t} (\alpha(j,t))^2, \end{aligned} \tag{9}$$

where $\alpha(j,t)$ is defined as $\sum_{i=j}^{t} \frac{\Theta\left(\frac{i+1}{t+1}\right)^{\lambda_{min}}}{i+1} \prod_{k=j+1}^{i} \gamma(k)$ and σ^2 is variance of channel noise. In preceding derivation, step (a) follows from equation (8). Step (b) is due to the fact that $v_1(i)$ and $v_2(i)$, the components of vector $\mathbf{v}(i)$, are uncorrelated random variables with identical distribution. Equation (9) completes the proof of part **(a)**.

Proof (b) Similar to equation (7), the communication energy can be written in terms of the initial values of local variables, the accumulated channel noise, and the optimum point where the latter is the dominant part. Therefore, the total communication energy for nodes S_1 and S_2 is

$$E_{com}(T) = \sum_{t=1}^{T} E\left\|\mathbf{x}(t) - \gamma(t)\mathbf{x}(t-1)\right\|^2$$
$$\geq \sum_{t=1}^{T}\sum_{l=1}^{2}(\bar{b_l})^2 \left(\left(1-\gamma(t)-\frac{\lambda_l}{t}\right)\sum_{i=1}^{t-2}\frac{\prod_{j=i+1}^{t-2}\left(1-\frac{\lambda_l}{j+1}\right)}{i+1}+\frac{1}{t}\right)^2$$

where $\bar{\mathbf{b}} := [\bar{b_1}, \bar{b_2}]^T = P^T\mathbf{b}$. In this equation, the inner summation can be written using equation (8):

$$\sum_{i=1}^{t-2}\frac{\prod_{j=i+1}^{t-2}\left(1-\frac{\lambda_l}{j+1}\right)}{i+1} = \frac{C_2\sum_{i=1}^{t-2}(i+1)^{\lambda_l-1}}{(t-1)^{\lambda_l}} = \Theta(1),$$

where C_2 is a positive constant. Therefore, we obtain

$$E_{com}(T) > \sum_{t=1}^{T}\sum_{l=1}^{2}(B_l)^2\left(1-\gamma(t)-\frac{D_l}{t}\right)^2 \qquad (10)$$

where D_l and $B_l, l \in \{1,2\}$, are positive constants. In part **(a)**, we proved that under condition (6), we have

$$\forall\ \epsilon > 0\ \ ,\exists\ t_\epsilon\ \ : E[\|\mathbf{e}(t_\epsilon+1)\|^2] \leq\ \epsilon. \qquad (11)$$

To prove part **(b)**, it is enough to show that for some constant C_3,

$$\epsilon\ E_{com}(t_\epsilon) \geq E\left[\|\mathbf{e}(t_\epsilon+1)\|^2\right]\ E_{com}(t_\epsilon) \geq C_3 > 0.$$

Using the Cauchy-Schwartz inequality and equations (9) and (11), we obtain

$$0 \leq \sum_{i=1}^{t_\epsilon}\frac{\alpha(i,t_\epsilon)}{i} \leq \sqrt{\sum_{i=1}^{t_\epsilon}\frac{1}{i^2}\sum_{i=1}^{t_\epsilon}(\alpha(i,t_\epsilon))^2} \leq C_4\sqrt{\epsilon}, \qquad (12)$$

where C_4 is a positive constant. It follows from equation (9) and (10) that

$$E_{com}(t_\epsilon)E\left[\|\mathbf{e}(t_\epsilon+1)\|^2\right]$$
$$\geq \sum_{l=1}^{2}(B_l)^2\sum_{i=1}^{t_\epsilon}\left(1-\gamma(i)-\frac{D_l}{i}\right)^2\sum_{i=1}^{t_\epsilon}(\alpha(i,t_\epsilon))^2$$
$$\overset{(a)}{\geq}\sum_{l=1}^{2}B_l\left(\sum_{i=1}^{t_\epsilon}\left(1-\gamma(i)-\frac{D_l}{i}\right)\alpha(i,t_\epsilon)\right)^2$$
$$\overset{(b)}{\geq}\sum_{l=1}^{2}B_l\left(\sum_{i=1}^{t_\epsilon}(1-\gamma(i))\sum_{j=i}^{t_\epsilon}\frac{\Theta\left(\frac{i+1}{t+1}\right)^{\lambda_{min}}}{i+1}\prod_{k=i+1}^{j}\gamma(k)-D_lC_4\sqrt{\epsilon}\right)^2$$
$$\overset{(c)}{=}\sum_{l=1}^{2}B_l\left(\sum_{i=1}^{t_\epsilon}\frac{\Theta\left(\frac{i+1}{t+1}\right)^{\lambda_{min}}}{i+1}-D_lC_4\sqrt{\epsilon}\right)^2$$
$$\geq\sum_{l=1}^{2}B_l\left(\Theta(1)-D_lC_4\sqrt{\epsilon}\right)^2\overset{(d)}{\geq}C_3.$$

In this equation, step (a) follows from the Cauchy-Schwartz inequality. Step (b) is due to equation (12), and definition of $\alpha(i,t_\epsilon)$. Step (c) is the result of changing the order of summation and step (d) holds for small enough ϵ.

3.1. Extension to the Multiple Node Case

Under some additional assumptions, Theorem **1** holds in the multiple node case. Assume that there exists a communication link between nodes S_i and S_j when $a_{i,j} \neq 0$. Furthermore, assume that nodes operation are synchronized. Then, the same proof of Section 3 applies here with minor changes in part **(a)** where the accumulated channel noise becomes

$$[\mathbf{v}(t)]_l = -\sum_{m=1,m\neq l}^{n}a_{l,m}\sum_{i=1}^{t}\prod_{k=i+1}^{t}\gamma(k)n^{m,l}(i)\ ,l\in\{1,\ldots,n\}.$$

Moreover, the same derivation of equation (9) applies and the mean square error can be written as

$$E[\|\mathbf{e}(t+1)\|^2] = \Theta\left(\frac{1}{t+1}\right)^{2\lambda_{min}} + \sum_{k,l=1,k\neq l}^{n}(a_{k,l})^2\sigma_{l,k}^2\sum_{i=1}^{t}(\alpha(i,t))^2.$$

4. OPTIMUM ENERGY DESIGNS

In this section, we prove that $\gamma^*(t) := exp(-t^{-q}), 0 < q < 0.5$ attains the minimum energy bound. For brevity, we omit the detail of derivations here. It follows from the definition of $\alpha(i,t)$ and Lemma **1** that:

$$\alpha(j,t) = \sum_{i=j}^{t}exp\left(-\sum_{k=j+1}^{i}\frac{1}{k^q}\right)$$
$$\leq\frac{1}{t}\left(1+exp\left(\frac{(j+1)^{1-q}}{1-q}\right)\int_{j+1}^{t+1}exp\left(-\frac{x^{1-q}}{1-q}\right)dx\right)\leq\frac{1+t^q}{t},$$

where the third step is due to the Holder's inequality. Since q is less than .5, the convergence condition in Theorem **1(a)** is satisfied. Moreover, the mean square error decreases asymptotically at the rate of t^{2q-1} (equation (9)). Similar to equations (7), (9), we can show the upper bound on communication energy grows at the rate of t^{2q-1} for $\gamma^*(t)$. This result along with Theorem **1(b)** prove that the communication energy required to compute ϵ-minimizer of $f(\mathbf{x})$ in mean square sense grows at the rate of $\Theta(\epsilon^{-1})$.

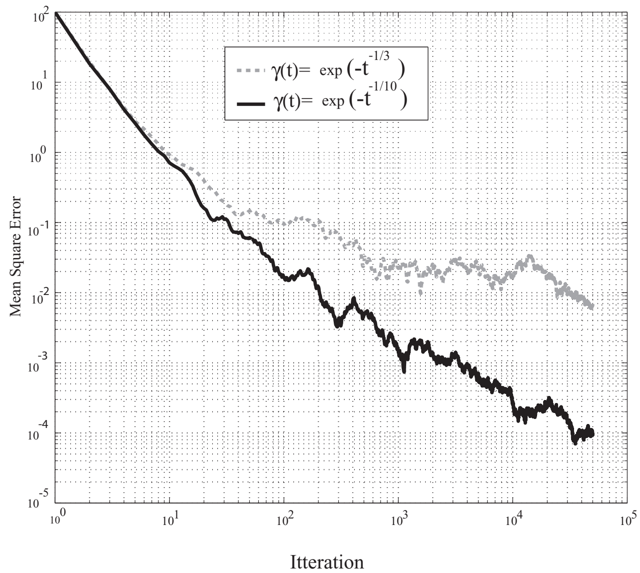

Fig. 1. Mean square error in x_1^1

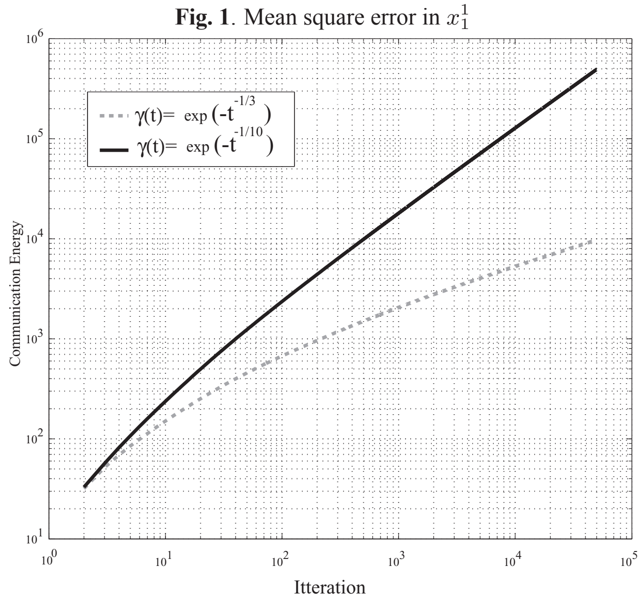

Fig. 2. The average of the communication energy

5. SIMULATION RESULTS

To illustrate the concept, a simple two dimensional example is considered. The cost function is a quadratic convex function with $A = \begin{bmatrix} 2.1 & 1 \\ 1 & 2.1 \end{bmatrix}$, $\mathbf{b} = [-11, 11]^T$, and a minimum point at $\mathbf{x} = [-10, 10]^T$; the channel noise is additive white gaussian noise with variance one. We also consider $exp(-t^{-q})$ for $q \in \{1/3, 1/10\}$ in simulations. Figure 1 shows the mean square error in x_1^1 averaged over 20 runs. This figure confirms that the mean square error decreases at the rate of t^{2q-1} when the accumulated channel noise becomes the dominant source of error. Therefore, a higher q results in a lower convergence rate as well as a lower increase of communication energy (Figure 2). Furthermore, the mean square error decreases linearly with the communication energy (Figure 3). In other words, that the communication energy required to compute ϵ-minimizer of $f(\mathbf{x})$ in mean square sense grows at the rate of $\Theta(\epsilon^{-1})$. These results agree with our theoretical analysis presented in pervious sections.

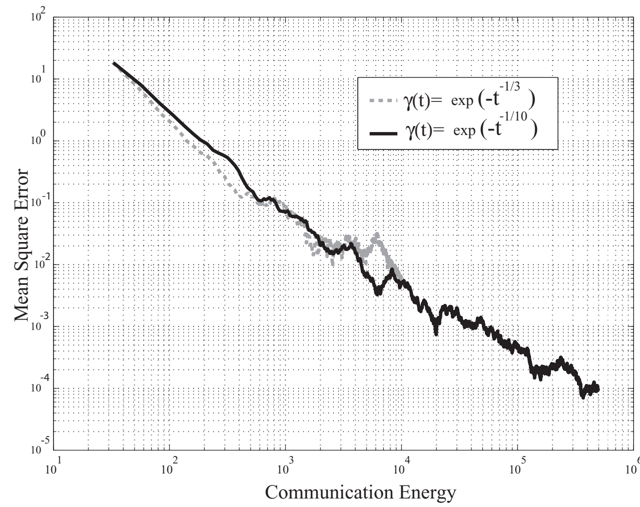

Fig. 3. Mean square error in x_1^1 versus the average of the communication energy

6. CONCLUSION AND FUTURE WORK

We studied the problem of distributed optimization of a general quadratic cost function in an energy-constrained network. We considered a class of distributed stochastic gradient type algorithms implemented using certain linear analog messaging schemes. It is shown that the communication energy to obtain an ϵ-minimizer of a cost function in the mean square sense must grow at least at the rate of $\Omega(\epsilon^{-1})$. We derived specific designs which attain the minimum energy bound, and confirmed our theoretical analysis by numerical simulations. The generalization of this work to a general non-quadratic cost function is possible and is a subject of ongoing research.

Acknowledgement : The authors gratefully acknowledge the assistance of Mikalai Kisialiou in the preparation of this manuscript.

7. REFERENCES

[1] D. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[2] M. W. Carter, H. H. Jin, M. A. Saunders, and Y. Ye, "Spaseloc: An adaptive subproblem algorithm for scalable wireless sensor network localization," *SIAM Journal on Optimization*, accepted March 2006.

[3] A. Olshevsky and J. N. Tsitsiklis, "Convergence speed in distributed consensus and averaging," in *45th IEEE Conference on Decision and Control*. San Diego, CA, USA: IEEE, December 2006.

[4] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[5] J. N. Tsitsiklis and Z. Luo, "Communication complexity of convex optimization," *Journal of Complexity*, vol. 3, no. 3, pp. 231–243, 1987.

[6] L. Xiao and S. Boyd, "Optimal scaling of a gradient method for distributed resource allocation," *To appear in Journal of Optimization Theory and Applications*, vol. 129, no. 3, pp. 469–488, 2006.