Stochastic Incremental Gradient Descent for Estimation in Sensor Networks

S. Sundhar Ram, A. Nedić and V. V. Veeravalli
University of Illinois at Urbana-Champaign
{ssriniv5,angelia,vvv}@uiuc.edu

Abstract—We consider a network of sensors deployed to sense a spatial field for the purposes of parameter estimation. Each sensor makes a sequence of measurements that is corrupted by noise. The estimation problem is to determine the value of a parameter that minimizes a cost that is a function of the measurements and the unknown parameter. The cost function is such that it can be written as the sum of functions (one corresponding to each sensor), each of which is associated with one sensor’s measurements. Such an objective function is of interest in regression. We are interested in solving the above optimization problem in a distributed and recursive manner. Towards this end, we combine the incremental gradient approach with the Robbins-Monro approximation algorithm to develop the Incremental Robbins-Monro Gradient (IRMG) algorithm. We investigate the convergence of the algorithm under a convexity assumption on the cost function and a stochastic model for the sensor measurements. In particular, we show that if the observations at each agent are independent and identically distributed, then the IRMG algorithm converges to the optimum solution almost surely as the number of observations goes to infinity. We emphasize that the IRMG algorithm itself requires no information about the stochastic model.

I. INTRODUCTION

Distributed computing has been a subject of extensive research in the last two decades. Initially, the purpose was to reduce the burden on a single processor and perform computationally large tasks in a distributed manner on multiple processors. Since the processors communicated through a wired medium, communication was usually reliable and not considered to be a scarce resource. With the recent interest in wireless sensor and actuator networks, the focus has shifted to performing simple global computations, e.g., averages, on data that is available locally at each agent in a distributed manner with minimum and unreliable communication. The recent literature on distributed computing in sensor networks solves the following three fundamental problems—optimization (e.g., [1]), function computation (e.g., [2]) and agent consensus (e.g., [3]). The algorithms for these fundamental distributed computation problems form the building blocks for distributed implementations of higher level operations such as detection, estimation, control, filtering and resource allocation. For example, the sum function computation is important in detection algorithms when each data point is the log-likelihood ratio. Another example is the consensus algorithm that is used in the distributed Kalman filter [3].

In this paper our focus is on distributed optimization. Specifically, we consider problems where the objective function can be decomposed into a sum of functions, each of which can be associated with a sensor’s measurements. Mathematically we are interested in solving the following optimization problem.

\[ x_n^* = \arg \min_{x \in X} \sum_{i=1}^{m} \frac{1}{n} \sum_{k=1}^{n} g_i(x, R_{i,k}). \]  

(1)

Here \( x \) is a vector parameter, \( X \) is the parameter space, and \( R_{i,k} \) is the measurement made by sensor \( i \) at time \( t_k \). Note that \( R_{i,k} \) could be a vector.

While there are many possible applications of problem (1) in sensor networks our primary motivation is least square regression. In particular, consider \( m \) sensors that are deployed to sense a spatial field. Let \( s \) denote the space index, and let \( s_i \) be the location of the \( i^{th} \) sensor. Consider a model set \( h(s, x) \), parameterized by \( x \), for the spatial field that the sensors are sensing. A model set is a set of candidate models that are decided based on \emph{a priori} information. Thus for each \( x \in X, h(s, x) \) is a model for the spatial field and \( h(s, x) \) is a model or predictor for the measurement \( R_{i,k} \). Note that the true model for \( R_{i,k} \) need not belong to the model set, i.e., the model set is just a ‘good guess’. The problem in regression is to use the measurements \( R_{i,k} \) to choose the best model in the model set, i.e., determine the value for \( x \), that best describes the spatial field. In least square regression the parameter value \( x_n^* \) is determined as follows.

\[ x_n^* = \arg \min_{x \in X} \sum_{i=1}^{m} \sum_{k=1}^{n} \frac{1}{n} (R_{i,k} - h(s_i, x))^2. \]  

(2)

If for some fixed value \( \hat{x} \in X \), the sensor measurements actually follow

\[ R_{i,k} = h(s_i, \hat{x}) + \epsilon_{i,k}, \]

where \( \epsilon_{i,k} \) can be modeled as i.i.d. zero-mean Gaussian noise then the \( x_n^* \) in (2) asymptotically converges to \( \hat{x} \) as \( n \to \infty \). If however \( R_{i,k} \) follows

\[ R_{i,k} = \hat{h}(s_i, \hat{x}) + \epsilon_{i,k}, \]

where \( \hat{h}(s_i, \hat{x}) \) is a function that is different from \( h(s_i, x) \), then \( x_n^* \) converges to

\[ x^* = \arg \min_{x \in X} \sum_{i=1}^{m} (h(s_i, x) - \hat{h}(s_i, \hat{x}))^2. \]
If the function \((h(s_i, x) - \tilde{h}(s_i, \tilde{x}))^2\) has multiple minima then \(x^*_n\) will converge to one of them.

We are interested in solving, or at least approximately solving, the problem in (1) in a distributed and recursive manner. By distributed we mean the a sensor should not share its data sequence with all the other agents. Equivalently, there is no central agent that collects all the observations.

By recursive we mean that the network should use obtain \(x^*_{n+1}\) directly from \(x^*_n\) and the new data \(R_{1,n+1}, \ldots, R_{m,n+1}\).

It is well known that except in some very specific cases a recursive solution does not exist in the centralized setting. In general, recursive approximations, \(x_n\), to the sequence \(x^*_n\) are generated that are asymptotically correct, specifically, as \(n \to \infty\), \(x_n\) converges to \(x^*\). For example, the least mean square (LMS) algorithm provides an asymptotically correct recursive solution to the regression problem when the function \(g_i(x, R_{i,k})\) is the squared error between \(R_{i,k}\) and \(x\).

We are interested in generating the iterates in a distributed and recursive manner. Therefore in generating the iterates two approximations would need to be made—one to obtain a recursive structure, and the other to obtain a distributed structure. Further the approximations should be such that the iterates converge asymptotically to \(x^*\). Our main contribution in this paper is to develop one such approximation scheme.

We combine the incremental gradient method \([4]\) with the Robbins-Monro stochastic approximation algorithm \([5]\) to develop a distributed and recursive solution to the optimization problem stated in (1).

We call our method the **Incremental Robbins-Monroe Gradient (IRMG)** algorithm. The iterate in the IRMG algorithm is passed in a cyclic manner from agent to agent. At each agent the received iterate is incrementally modified and passed on to the next agent. See Fig. 1 for an illustration. As in the standard Robbins-Monro algorithm \([5]\) (also see Sec. II-B), we study the convergence properties of the IRMG algorithm under a stochastic model for the observations. In particular, we model \(\{R_{i,k}\}_{k=1,2,\ldots}\) as independent samples of a random variable \(R_i\). Under this stochastic model for the observations, it is easily seen that optimization problem of (1) takes on the following limiting form as \(n \to \infty\).

\[
x^* = \arg \min_{x \in X} \sum_{i=1}^{m} E[g_i(x, R_i)].
\]

We show that the iterates generated by the IRMG algorithm converge to \(x^*\) almost surely as long the functions \(g_i(x, R_i)\) are convex in \(x\).

We emphasize that the actual algorithm requires no information about the statistics of \(R_i\), or even the fact that the measurements are i.i.d. samples of \(R_i\). We need the stochastic model for \(R_{i,n}\) only to establish convergence.

The rest of the paper is organized as follows. In Section II, we discuss some important preliminaries that we use in the development of our algorithm. In Section III, we develop the IRMG algorithm and discuss its convergence properties. In Section IV, we verify our theoretical results through simulation for an illustrative example. Finally, in Section V, we provide some concluding remarks.

**II. Preliminaries**

In this section we briefly describe the two main ingredients of our algorithm—the incremental gradient method and the Robbins-Monro stochastic approximation algorithm.

**A. Incremental Gradient Method**

The incremental gradient method (\([4]\), and references therein) is an algorithm to minimize a sum of convex functions in a distributed manner. In particular, suppose we are interested in minimizing

\[
f(x) = \sum_{i=1}^{m} f_i(x), \quad x \in X.
\]

where \(f_i\) is convex for each \(i\).

The standard gradient method for this minimization problem would compute iterates in time \((n)\) as follows:

\[
x_{n+1} = P_X [x_n - \alpha_n \nabla f(x_n)].
\]

Here \(\nabla\) denotes the gradient, \(P_X\) denotes projection onto the set \(X\), and \(\alpha_n\) is the step size.

The incremental gradient method is similar to the standard gradient descent method; however, each step in this method consists of \(m\) sub-steps where the iterate is incrementally changed and communicated from one sensor to the next. Hence the name incremental gradient. Mathematically, the recursion for the algorithm as a function of time can be described by the following equations:

\[
x_n = z_{m,n} = z_{0,n+1}
\]

\[
z_{i,n+1} = P_X [z_{i-1,n+1} - \alpha_n \nabla f_i(z_{i-1,n+1})].
\]

The algorithm can be understood as follows. Associate with sensor \((i-1)\) the function \(f_{i-1}()\) and view \(z_{i,n+1}\) as the iterate received by sensor \(i\) at time \(n+1\). Sensor \(i\) then incrementally modifies the iterate using the gradient of ‘its function’, i.e., \(f_i()\), evaluated at the received iterate, \(z_{i-1,n+1}\), and passes the updated iterate to the next sensor. See Fig. 1 for an illustration. Thus the key difference between the gradient and incremental gradient method is that the gradient of \(f_i()\) at iteration \(n+1\) is evaluated at \(x_n\) in the standard gradient method, while it is evaluated at \(z_{i-1,n+1}\) in the incremental gradient method. The convergence of the incremental gradient method is studied in \([4]\).

There are two main benefits to the incremental gradient method. First, the incremental gradient method requires less
communication than the standard gradient method since the sensors do not need to send their observations to a central
processor (fusion center) at each time step for optimization [1]. Second, the incremental gradient method converges faster
than the standard gradient method when the iterates are far from the eventual limit [6].

B. Robbins-Monro gradient algorithm

The Robbins-Monro gradient algorithm is based on the idea of stochastic approximation introduced in the seminal paper by Robbins and Monro [5]. Suppose we are interested in the following minimization problem:

$$x^* = \arg \min_{x \in X} \frac{1}{n} \sum_{k=1}^{n} g(x, R_k).$$

(6)

The Robbins-Monro algorithm provides the following recursive approximation to $$x^*_n$$. For any $$n$$, given the approximation $$x_n$$ to $$x^*_n$$, we can form the approximation $$x_{n+1}$$ for $$x^*_n$$ as

$$x_{n+1} = \mathcal{P}_X \left[ x_n - \alpha_n \nabla g(x_n, R_{n+1}) \right].$$

The idea is to use the empirical gradient $$\nabla g(x_n, R_{n+1})$$ at time $$n$$ instead of the actual gradient

$$\frac{1}{n} \sum_{k=1}^{n} \nabla g(x_n, R_k).$$

The algorithm is clearly recursive.

Now suppose that the measurement $$R_n$$ can be modeled as independent samples of a random variable $$R$$. Then, the optimization problem of (6) takes on the following limiting form as $$n \to \infty$$:

$$x^* = \arg \min_{x \in X} E[g(x, R)].$$

It is shown in [5] that the recursive approximation $$x_n$$ converges to $$x^*$$ almost surely as $$n \to \infty$$, i.e., that the Robbins-
Monro procedure is asymptotically correct.

III. INCREMENTAL ROBBINS-MONRO GRADIENT METHOD

We combine the incremental gradient method and the Robbins-Monro algorithm to develop the Incremental
Robbins-Monro Gradient (IRMG) algorithm. In the IRMG
algorithm, the iterates of the approximation to the solution to (1) are obtained according to

$$x_n = z_{m,n} = z_{0,n+1}$$

$$z_{i,n+1} = \mathcal{P}_X \left[ z_{i-1,n+1} - \alpha_n \nabla g_i(R_{i,n+1}, z_{i-1,n+1}) \right]$$

(7)

The idea is as follows. We use the empirical Robbins-Monro gradient instead of the actual gradient of $$f_i(x)$$ in the incremental
gradient method. Observe that the algorithm has a natural distributed and recursive structure. The iterate $$x_{n+1}$$ is obtained from $$x_n$$ only, along with the data $$R_{1,n+1}, \ldots, R_{m,n+1}$$. Therefore it is recursive. In each such iteration each sensor needs to only pass the iterate around in a cycle.

We now state the main theorem of the paper.

Theorem 1: Let $$x^*$$ denote the limit of the sequence $$\{x_n^*\}$$ defined in (1). The iterates generated by the IRMG algorithm

in (7) converge to $$x^*$$ almost surely if the following conditions hold:

1) The set $$X$$ is nonempty, closed and convex.
2) The function $$f_i(x)$$ is convex in $$x$$.
3) For each $$i$$, $$\{R_{i,n}\}_{n=1}^{\infty}$$ are i.i.d samples of a random variable $$R_i$$, where $$R_i$$ satisfies $$E[\|\nabla g_i(x, R_i)\|^2] \leq K_i < \infty$$ for all $$x \in X$$. Further, the samples $$R_{i,n}$$ are independent across sensors also.
4) The step size satisfies

$$\sum_n \alpha_n = \infty \quad \text{and} \quad \sum_n \alpha_n^2 < \infty$$

We make the following remarks about the result. Condition 1 is a regularity condition that requires the moments of the measurements to be bounded. Condition 2 restricts the class of functions where the algorithm is asymptotically correct to only convex functions. However in a separate study we have extended the result to some non-convex functions that are important in the context of regression [7]. Condition 3 requires the measurements to be independent across both time and sensors. It is possible to relax this assumption and assume independence only across time by making the stronger convexity assumption that $$g_i(R_i, x)$$ is convex in $$x$$ for all $$R_i$$. Finally the last condition restricts the step-size to a diminishing step-size.

We next study the rate of convergence of the IRMG algorithm. Let $$\text{dist}(x_n, x^*)$$ denote the distance between $$x_n$$ and $$x^*$$. Theorem 2: Let the assumption made in Theorem 1 hold. Further, let the function $$f(x)$$ have a sharp set of minima, i.e.,

$$f(x) - f(x^*) \geq \mu (\text{dist}(x))^2 \quad \forall \quad x \in X.$$ 

Fix $$\alpha_n = \frac{1}{n^s}$$ for $$\frac{1}{2} < q \leq 1$$ and $$c = 2\mu A$$. Then

$$E[\text{dist}(x_n, x^*)] \leq O \left( \frac{1}{n^q} \right) \quad \text{if } 0 < q < 1$$

$$E[\text{dist}(x_n, x^*)] \leq O \left( \frac{1}{n} \right) \quad \text{if } q = 1 \text{ and } c > 1$$

$$E[\text{dist}(x_n, x^*)] \leq O \left( \frac{\log n}{n} \right) \quad \text{if } q = 1 \text{ and } c = 1$$

$$E[\text{dist}(x_n, x^*)] \leq O \left( \frac{1}{n^c} \right) \quad \text{if } s = 1 \text{ and } c < 1.$$ 

The function $$f(x)$$ has a sharp set of minima if it strongly convex, or the set $$X$$ is a polyhedral. We also mention that the result only provides an upper bound on the rate of convergence. The iterates could actually converge at a faster rate. The proofs are available in Appendix A.

IV. NUMERICAL SIMULATIONS

We consider a spatially deployed network of 200 sensors used for measuring the average pollution level over a region. The actual pollution level has a spatial and temporal variation. The objective is to determine a single pollution value that best approximates the spatio-temporal field. Such a concise representation of the field may be sufficient for some applications.

The sensors are indexed as $$1, \ldots, 200$$, in some order. We use $$p_{i,k}$$ to denote the actual pollution level at the $$i^{th}$$ sensor location at the instant the $$k^{th}$$ sample is taken. We use $$R_{i,k}$$
to denote the measurement made by the \(i\)th sensor in the \(k\)th sampling slot. The measurements are noisy, i.e.,

\[ R_{i,k} = p_{i,k} + n_{i,k}, \]

where \(n_{i,k}\) is the measurement noise. In robust least square regression, \([8]\), the estimate of the average pollution level is given by

\[ x_n^* = \arg \min_{x \geq 0} \sum_{i=1}^{200} \frac{1}{n} \sum_{k=1}^{n} g(x - R_{i,k}), \tag{8} \]

where

\[ g(x) = \log \left( \frac{|x|}{c} + 1 \right) - \log \left( \frac{|x|}{c} \right). \]

Note that the function \(g(x)\) is strictly convex. The data is generated such that the conditions in Theorem 1 are satisfied.

The actual pollution levels \(p_{i,k}\) are generated as i.i.d. instances of a random variable that is uniformly distributed in the range \([7, 13]\). The measurements noise sequence \(\{n_{i,k}\}\) is generated as independent samples of a zero-mean Gaussian with variance 1, for normal sensors, and variance 10, for faulty sensors. Thus we expect \(x_n^*\) in (8) to converge to

\[ x^* = \arg \min_{x \geq 0} \sum_{i=1}^{200} \mathbb{E}[g(x - R_{i,k})]. \tag{9} \]

We run 100 iterations of centralized Robbins-Monro gradient algorithm and the IRMG algorithm to minimize (9). The step size used for both algorithms is \(\alpha_n = \frac{a}{n}\). A comparison of the algorithms is presented in Figure 2. First observe that iterates generated by the two algorithms converge to the same values in about 50 iterations. This validates the convergence result of Theorem 1. Further, note that the incremental algorithm converges faster than the centralized version. This behavior can be attributed to taking projections at every sub-step in the incremental algorithms.

In practice the pollution level can be modeled as identically distributed random variables only over a few hours duration. For example, the pollution level in the evening is significantly different from the pollution level in the early mornings. If we assume that each sensor can sample every 30 seconds, then for the measurement variance and step-size used, the algorithm seems to converge in 50 iterations or about 25 minutes. Thus by running separate runs of the algorithm over intervals of a few hours we can track the variation in the average pollution level.

V. CONCLUDING REMARKS

In this paper we obtained convergence results for an incremental version of the Robbins-Monro stochastic approximation algorithm, the IRMG algorithm, under diminishing step-sizes. The IRMG algorithm provides a distributed and recursive solution to an optimization problem that arises in regression in sensor networks. We remark that the incremental LMS algorithm proposed in \([9]\) for linear regression is a specific case of the general problem we study. We also have results for a constant step-size IRMG algorithm that we could not report here. We also have similar results when the Kiefer-Wolfowitz empirical gradient \([10]\) is used in place of the Robbins-Monro gradient. In an another study, \([7]\), we are investigating the case when \(R_{i,k}\) is not an i.i.d. sequence. Finally, a problem of future interest is the convergence behavior of the incremental algorithms in an asynchronous communication setting with dynamic topologies, similar to the models in \([11]\).

REFERENCES


APPENDIX

A. Proof of Theorem 1

The analysis is based on the following theorem due to Robbins and Siegmund \([12]\).
Theorem 3: Let \{A_k\}, \{B_k\} and \{C_k\} be non-negative random sequences. Define \(F_k\) the be \(\sigma\)-algebra generated by \(A_1, \ldots, A_k, B_1, \ldots, B_k, C_1, \ldots, C_k\). If \(p\) w.p.1, \(\sum_k C_k < \infty\) and \(\mathbb{E}[A_{k+1} | F_k] \leq A_k - B_k + C_k\), then \(A_k\) converges to a non-negative quantity w.p.1 and \(\sum_k B_k < \infty\) w.p.1.

Next we state two inequalities from [12] that are a consequence of the convexity of the functions \(f_i\).

\[
f_i(x_1) - f_i(x_2) \geq (\nabla f_i(x_2))^T (x_1 - x_2), \quad (10)
\]

\[
f_i(x_1) - f_i(x_2) \leq (\nabla f_i(x_1))^T (x_1 - x_2). \quad (11)
\]

If \(f_i\) is not differentiable, \(\nabla f_i\) is to be interpreted as the subgradient. We will also use the following notation:

\[
z_{i,k+1} - x^* = d_{i,k+1}, \quad \text{and} \quad x_{k+1} - x^* = d_{k+1}.
\]

We first establish some inequalities.

\[
||d_{i,k+1}||^2 = ||P_X[z_{i-1,k+1} - \alpha_i \nabla g(R_{i-1,k+1}, z_{i-1,k+1})] - x^*||^2 \\
\leq ||z_{i-1,k+1} - \alpha_i \nabla g(R_{i-1,k+1}, z_{i-1,k+1}) - x^*||^2 \\
= ||d_{i-1,k+1}||^2 - 2\alpha_i d_{i-1,k+1}^T \nabla g(R_{i-1,k+1}, z_{i-1,k+1}) + \alpha_i^2 ||\nabla g(R_{i-1,k+1}, z_{i-1,k+1})||^2. \quad (12)
\]

The first inequality follows from the non-expansive property of the Euclidean projection function on closed convex sets. Define \(F_{i,k}^{-1}\) to be the \(\sigma\)-algebra generated by sample values \(R_{1,1}, \ldots, R_{M,1}, \ldots, R_{i-1,1}\). Note that the iterates \(z_{1,1}, \ldots, z_{i-1,k+1}\) also belong to the \(\sigma\)-algebra. Taking conditional expectations in (12)

\[
\mathbb{E}[||d_{i,k+1}||^2 | F_{i,k}^{-1}] \\
\leq ||d_{i-1,k+1}||^2 + \alpha_i^2 \mathbb{E}[||\nabla g_i(R_{i-1,k+1}, z_{i-1,k+1})||^2 | F_{i,k}^{-1}] \\
- 2\alpha_i d_{i-1,k+1}^T \mathbb{E}[\nabla g_i(R_{i-1,k+1}, z_{i-1,k+1}) | F_{i,k}^{-1}] \\
\leq ||d_{i-1,k+1}||^2 + 2\alpha_i K_i \\
- 2\alpha_i d_{i-1,k+1}^T \mathbb{E}[\nabla g_i(R_{i-1,k+1}, z_{i-1,k+1}) | F_{i,k}^{-1}]. \quad (13)
\]

The last step follows as the moments of \(R_i\) are bounded (Condition 3 in Theorem 1). Note that \(\mathbb{E}[\nabla g_i(R_{i-1,k+1}, z_{i-1,k+1}) | F_{i,k}^{-1}]\) can be written as \(\mathbb{E}[\nabla g_i(R_{i-1,k+1}, z_{i-1,k+1}) | z_{i-1,k-1}]\). This is because the samples are independent across both time and sensors. The quantity \(\mathbb{E}[\nabla g_i(R_{i-1,k+1}, z_{i-1,k+1}) | z_{i-1,k-1}]\) is simply \(\nabla f_i(x)\) evaluated at \(x = z_{i-1,k-1}\). We will denote this quantity by \(\nabla f_i(z_{i-1,k-1})\). We next bound the second term. Since the function \(f_i(x)\) is convex, from (11)

\[
d_{i-1,k+1}^T \mathbb{E}[\nabla g_i(R_{i-1,k+1}, z_{i-1,k+1}) | F_{i,k}^{-1}] \\
\geq f_i(z_{i-1,k-1}) - f(x^*) \\
= (f_i(x_k) - f_i(x^*)) + (f_i(z_{i-1,k-1}) - f_i(x_k)) \\
\geq (f_i(x_k) - f_i(x^*)) + \|\nabla f_i(x_k)^T (z_{i-1,k-1} - x_k)\| \\
\geq (f_i(x_k) - f_i(x^*)) - \|\nabla f_i(x_k)(z_{i-1,k-1} - x_k)\|. \quad (14)
\]

Equation (14) follows from (10). From Jensens’ inequality

\[
\|\nabla f_i(x_k)\| = \|\mathbb{E}[g_i(R_i, x_k) | x_k]\| \leq \sqrt{\mathbb{E}[\|g_i(R_i, x_k) | x_k\|^2]}.
\]

Using this,

\[
d_{i-1,k+1}^T \mathbb{E}[\nabla g_i(R_{i-1,k+1}, z_{i-1,k-1}) | F_{i,k}^{-1}] \\
\geq (f_i(x_k) - f_i(x^*)) - \sqrt{K_i} \|z_{i-1,k-1} - x_k\|.
\]

Therefore, we can strengthen the inequality in (13)

\[
\mathbb{E}[||d_{i,k+1}||^2 | F_{i,k}^{-1}] \leq ||d_{i-1,k+1}||^2 + \alpha_i^2 K_i \\
- 2\alpha_i (f_i(x_k) - f_i(x^*)) + 2\alpha_i \sqrt{K_i} \|z_{i-1,k+1} - x_k\|. \quad (15)
\]

Next we will bound \(\mathbb{E}[\|z_{i-1,k+1} - x_k \| | F_{i,k}^M]\).

\[
\|z_{i-1,k+1} - x_k\| \leq \sum_{j=1}^{i-1} \|z_{j,k+1} - z_{j-1,k+1}\| \\
\leq \sum_{j=1}^{i-1} \alpha_k \|\nabla g_i(R_{j,k+1}, z_{j-1,k+1})\|.
\]

Therefore,

\[
\mathbb{E}[\|z_{i-1,k+1} - x_k \| | F_{i,k}^M] \leq \sum_{j=1}^{i-2} \alpha_k \|\nabla g_i(R_{j,k+1}, z_{j-1,k+1})\| + \alpha_k \sqrt{K_i}. \quad (16)
\]

Taking expectations in a step-by-step manner

\[
\mathbb{E}[\|z_{i-1,k+1} - x_k \| | F_{i,k}^M] \leq \alpha_k \sum_{j=1}^{i-1} \sqrt{K_j}.
\]

Taking expectations in (15) and using this inequality

\[
\mathbb{E}[||d_{i,k+1}||^2 | F_{i,k}^M] \leq \mathbb{E}[||d_{i-1,k+1}||^2 | F_{i,k}^M] + \alpha_i^2 K_i \\
- 2\alpha_i (f_i(x_k) - f_i(x^*)) + 2\alpha_i^2 K_i \sum_{j=1}^{i-1} \sqrt{K_j}.
\]

Summing over all \(i\) we get

\[
\mathbb{E}[||d_{k+1}||^2 | F_{k}^M] \leq \|d_k\|^2 \\
+ \alpha_k^2 \sum_{i=1}^{M} K_i - 2\alpha_k (f_i(x_k) - f_i(x^*)) + 2\alpha_k^2 \sum_{i=1}^{M} K_i \sum_{j=1}^{i-1} \sqrt{K_j} \\
\leq \|d_k\|^2 + \alpha_k^2 \left( \sum_{i=1}^{M} \sqrt{K_i} \right)^2 - 2\alpha_k (f_i(x_k) - f_i(x^*)). \quad (16)
\]

Observe that the equation now satisfies the conditions of Theorem 3. Therefore, we can conclude w.p.1. that \(\|d_k\|\) converges and \(\sum_k \alpha_k (f_i(x_k) - f_i(x^*)) < \infty\). Since \(\sum_k \alpha_k = \infty\) it follows that \(f_i(x_k) \to f_i(x^*)\). Using sample path arguments we can now argue that \(x_k \to x^*\).

B. Proof of Theorem 2

Using the strict inequality assumption in (16) we get

\[
\mathbb{E}[||d_{k+1}||^2 | F_{k}^M] \leq (1 - 2\alpha_k \mu) \|d_k\|^2 + \alpha_k^2 \left( \sum_{i=1}^{M} \sqrt{K_i} \right)^2.
\]

Taking expectations in

\[
\mathbb{E}[||d_{k+1}||^2] \leq (1 - 2\alpha_k \mu) \mathbb{E}[\|d_k\|^2] + \alpha_k^2 \left( \sum_{i=1}^{M} \sqrt{K_i} \right)^2.
\]

The result now follows from Lemma 4, Page 45 in [12].