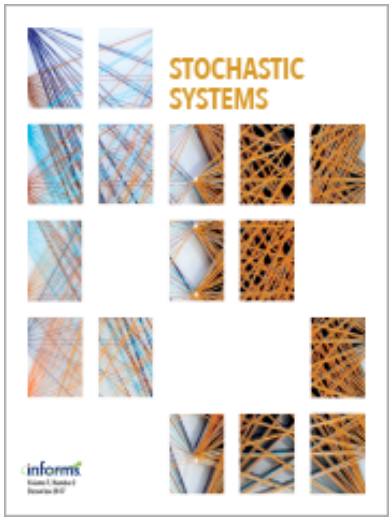

Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Convergence of a Distributed Kiefer-Wolfowitz Algorithm

Jean Walrand

To cite this article:
Jean Walrand (2021) Convergence of a Distributed Kiefer-Wolfowitz Algorithm. Stochastic Systems 11(4):324-332.
https://doi.org/10.1287/stsy.2021.0080

https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 11, No. 4, December 2021, pp. 324–332
ISSN 1946-5238 (online)

Convergence of a Distributed Kiefer-Wolfowitz Algorithm

Jean Walrand[a]

[a] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, California 94720
Contact: walrand@berkeley.edu, https://orcid.org/0000-0002-1460-6826 (JW)

Received: August 28, 2020
Revised: December 21, 2020
Accepted: March 15, 2021
Published Online in Articles in Advance: September 13, 2021

https://doi.org/10.1287/stsy.2021.0080

Abstract. This paper proposes a proof of the convergence of a distributed and asynchronous version of the Kiefer-Wolfowitz algorithm where the agents do not exchange information with one another.

Keywords: stochastic approximation • stochastic gradient descent • distributed • asynchronous

1. Introduction

The goal is to maximize a smooth concave function of $K > 1$ variables. The function is assumed to be strongly concave in a neighborhood of its maximizer. There are K agents; each agent observes the values of the function, corrupted by observation noise, and adjusts its own variable without knowing the values of the other variables. The agents do not communicate their variables. This formulation is motivated by many applications where the agents do not know each other or are not be able to communicate directly with one another. Moreover, the agents are not synchronized, so that they update their variable either at the same or different times.

Each agent *experiments* by perturbing its variable by a zero-mean change in order to estimate the partial derivative of the function with respect to that variable. In the next step, the agent then *updates* its variable in proportion to the estimate of the partial derivative. Each agent experiments exactly once in every interval of τ successive steps, for some common $\tau \geq 2$. A particular example is when the agents update with the same period τ but may be out of phase. This is admittedly a restricted form of asynchronicity.

This algorithm is an extension of Kiefer and Wolfowitz (1952) and Spall (1992). In Kiefer and Wolfowitz (1952), the authors introduce a gradient descent algorithm where the gradient is estimated by observing the function at perturbed values of its variable and they prove the convergence of the algorithm to the minimum of the function. Spall (1992) proposes a variation of the algorithm in the multivariate case where the partial derivatives with respect to the different variables are estimated by simultaneously perturbing each variable by an independent and zero-mean amount, instead of perturbing the variables one at a time. The author proves the convergence to the minimum of the function under the assumption that the variables return infinitely often to a compact set. In this paper, we extend the algorithm to the case where the different variables get updated asynchronously. Also, the proof does not require assuming returns to a compact set. This assumptions is needed in proofs based on the ordinary differential equation approach, such as in Kushner and Clark (1978) and Ljung and Söderström (1983). The key idea in this approach is to show that the piecewise linear interpolation of the successive values of the vector of variables, when situated at appropriately scaled time steps and restarted in a compact neighborhood of the maximizer of the function, approaches the solution of a differential equation that converges to that maximizer.

An agent corrupts the estimate of the partial derivative of another agent either when it experiments or updates its variable while the other agent calculates its estimate. Technically, the difficult aspect of our version is that the corruption of the estimate by the updates of other agents is not zero-mean, in contrast with the corruption by their experiments, which is zero-mean (up to a second order term). Intuitively, the result is not surprising; but its formal proof requires careful bounds on the size of the corruptions. This is the technical contribution of the paper.

Some papers propose mechanisms where agents exchange the value of their variables, possibly with some delays, and they may know the function they want to maximize (e.g., Nedic et Ozdaglar 2009, Kennedy et al. 2019, Ramaswamy 2019, Swenson et al. 2020). The key contribution of this paper is to show that such communication is

not necessary for convergence. Also, the agents observe the values of the function with some observation noise but need not know its functional form. That is, the agents can observe the effect of their choice of value for their variable, but they could not calculate it.

2. Algorithm and Result

Recall that there is some integer $\tau \geq 2$ such that, for each $n \geq 0$, every agent k experiments exactly once at some time $T_k(n)+1$, where $T_k(n) \in \{n\tau, \ldots, n\tau+\tau-1\}$, and updates at the next step. Let $f : \Re^K \to \Re$ be some function defined on $\Re^K$. The case of a single agent (i.e., $K=1$) is the same as in Kiefer and Wolfowitz (1952), whereas that of simultaneous updates (i.e., $\tau = 1$) corresponds to Spall (1992).

The experiments and updates are defined as follows. For $k = 1, \ldots, K$ and $n \geq 0$, let $x_k(n)$ be the value of the variable of agent k at step n. Let also $\mathbf{x}(n)$ be the vector with components $x_k(n)$, for $n \geq 0$.

The algorithm is as follows. For $k = 1, \ldots, K$ and $n \geq 0$, one has, for $m = T_k(n)$,

$$x_k(m+1) = x_k(m) + a_k(n)\epsilon(n) \quad \text{(experiment)}, \tag{1}$$

$$x_k(m+2) = x_k(m) + g_k(n)\gamma(n) \quad \text{(update)}, \tag{2}$$

where

$$g_k(n) = \frac{f(\mathbf{x}(m+1)) - f(\mathbf{x}(m)) + \eta_k(n)}{x_k(m+1) - x_k(m)}; \tag{3}$$

$$a_k(n) \text{ are independent with } P(a_k(n) = -1) = P(a_k(n) = 1) = 0.5; \tag{4}$$

$$\eta_k(n) \text{ are independent, zero} - \text{mean, bounded}; \tag{5}$$

$$\epsilon(n), \gamma(n) \in (0,1), \epsilon(n) \to 0, \sum \gamma(n) = \infty, \sum \frac{\gamma^2(n)}{\epsilon^2(n)} < \infty; \tag{6}$$

$$\sum \gamma(n)\epsilon^2(n) < \infty, \gamma(n)/\epsilon^2(n) \text{ is bounded}. \tag{7}$$

(For instance, $\gamma(n) = n^{-0.75}, \epsilon(n) = n^{-0.2}$.)

Also, if $m - 1 \neq T_k(n)$ and $m \neq T_k(n)$ for any n, then

$$x_k(m+1) = x_k(m). \tag{8}$$

The assumptions on the step size $\epsilon(n)$ and the experimentation size $\gamma(n)$ are similar to those of Kiefer and Wolfowitz (1952) and Spall (1992). Specifically, these two references require (6). Kiefer and Wolfowitz (1952) requires $\sum \gamma(n)\epsilon(n) < \infty$ instead of (7). In particular, if $\gamma(n) = n^{-a}$ and $\epsilon(n) = n^{-b}$, then the conditions in Kiefer and Wolfowitz (1952) subsume our conditions; also, $a = 0.75, b = 0.1$ satisfy the conditions in Spall (1992) but not ours. We assume that the noise is bounded, whereas Kiefer and Wolfowitz (1952) and Spall (1992) only impose conditions on moments. The proof techniques are different.

Our objective is to prove the following theorem.

Theorem 1. *Assume that the function f is smooth and concave, with a finite maximizer $\mathbf{x}^*$, strongly concave in a neighborhood of $\mathbf{x}^*$, and with bounded derivatives up to the third order. Then the algorithm defined by (1)–(8) is such that*

$$\mathbf{x}_n \to \mathbf{x}^*, \text{ almost surely as } n \to \infty.$$

3. Convergence Rate

Assume that $\gamma(n) = n^{-a}$ and $\epsilon(n) = n^{-b}$. Our conditions require $a > \max\{2b, 1-2b, 0.5+b\}$. Assume further that $a < 6b$. (For instance, $a = 0.75$ and $b = 0.2$.) Then Spall (1992) shows that the synchronized version of the algorithm ($\tau = 1$) is such that, for $n \gg 1$,

$$n^{a/2-b}(\mathbf{x}_n - \mathbf{x}^*) \approx \mathcal{N}(0, \Sigma),$$

where Σ is a matrix that depends on the Hessian of f at $\mathbf{x}^*$.

In the asynchronous version of the algorithms, all the variables get updated every $\tau+1$ steps instead of every two steps. Also, the errors in the gradient estimation are shown to be second order in the proof of the theorem. Thus, one expects the convergence to occur $(\tau+1)/2$ times more slowly than for the synchronized algorithm. The confirmation of this estimate is left for further study.

4. Outline of Proof of Theorem

Let $\mathbf{z}(m)$ be the vector with components $z_k(m) = x_k(m) - a_k(n)\epsilon(n)1\{m = T_k(n)+1\}$. That is, $z_k(m)$ is the latest updated value of x_k by time m. Note that $z_k(m)$ does not change when user k performs an experiment, only when it updates its variable. Of course, $\mathbf{x}$ changes during experiments and updates; and the gradient is estimated by observing $f(\mathbf{x})$, not $f(\mathbf{z})$. Let also L be a bound on the first derivative of f.

Fix any $\delta > 0$. It is shown in the next section that $u(n) = \|\mathbf{z}(n\tau) - \mathbf{x}^*\|$ satisfies the following two inequalities:

$$u(n+1) \le u(n) - [\gamma(n)\beta - \alpha(n)], \text{ whenever } u(n) > \delta \tag{9}$$

and

$$u^2(n+1) \le u^2(n) + c(n), \text{ whenever } u(n) \le \delta. \tag{10}$$

In these expressions, $\beta > 0, c(n) \to 0,$ and $\sum\alpha(n)$ converge to a finite random variable.

The claim is that these inequalities imply that $u(n) \le 3\delta$ for all $n \ge n_0$ for some finite n_0. To see this, choose n_0 so that $c(n) \le 3\delta^2$ for $n \ge n_0 - 1$ and $\sum_{n=n_0}^{n_0+m}\alpha(n) \le \delta$ for all $m \ge 0$. Let n_1 be the first time after n_0 that $u(n) > \delta$. If there is no such time, we are done. Else, let m_1 be the first time after n_1 that $u(n) \le \delta$. Such a time must exist because of (9), for otherwise $u(n) \to -\infty$ because $\sum\gamma(n) = \infty$ and $\sum\alpha(n) < \infty$. Let then n_2 be the first time after m_1 that $u(n) > \delta$, then m_2 the first time after n_2 that $u(n) \le \delta$, and so on. Finally, let $v(j)$ be the maximum value of $u(n)$ for $n \in \{n_j, \ldots, m_j - 1\}$. Because of (10), $u^2(n_j) \le u^2(n_j - 1) + c(n_j - 1) \le \delta^2 + 3\delta^2$, so that $u(n_j) \le 2\delta$. Also, because of (9), $v(j) - u(n_j) \le \max_m \sum_{n=0}^{n_0+m}\alpha(n) \le \delta$. Hence, $v(j) \le 3\delta$ for all j, so that $u(n) \le 3\delta$ for all $n \ge n_0$.

Because $\delta > 0$ is arbitrary, it follows that $u(n) \to 0$. Because $\|\mathbf{z}(n\tau) - \mathbf{x}(n\tau)\| \le \epsilon(n)$, this implies that $\mathbf{x}(n) \to \mathbf{x}^*$.

5. Proof of (9) and (10)

We first give the main steps that lead to the inequalities. The rest of the section provides the details of the calculations.

5.1. Main Steps

Inequality (9) says that when $\mathbf{z}$ is away from the maximizer $\mathbf{x}^*$, the gradient updates bring it closer. This is intuitive because the gradient is then large. Inequality (10) says that when $\mathbf{z}$ is close to the maximizer, the updates do not make it move faraway. This happens because the gradient is then small.

Every τ steps, each variable x_k gets updated roughly in the direction of the partial derivative of $f(\cdot)$ with respect to that variable. Thus, $\mathbf{z}$ gets updated roughly in the direction of the gradient $\nabla f(\mathbf{x})$. Errors occur because of corruptions of the gradient estimate due to observation noise and the changes of the other variables by other agents. More precisely, using (3) one finds (see Lemma 4)

$$\begin{aligned}\mathbf{w}(n) &:= \mathbf{z}(n\tau + \tau) - \mathbf{z}(n\tau)\\ &= \gamma(n)\nabla f(\mathbf{z}(n\tau)) + \mu(n)\gamma(n)/\epsilon(n) + O(\rho(n)),\end{aligned} \tag{11}$$

where $\rho(n) = \max\{\gamma^2(n)/\epsilon(n), \gamma(n)\epsilon^2(n)\}$ and $\mu(n)$ are a bounded random vector that is zero-mean given $\mathcal{F}_{n-1}$, where

$$\mathcal{F}_n := \{a_k(m), \eta_k(m), m \le n; k = 1, \ldots, K\}.$$

Also, in (11), $O(\rho(n))$ is a random vector whose components are bounded in absolute value by a constant times $\rho(n)$.

Identity (11) implies (see Lemma 5)

$$\|\mathbf{w}(n)\|^2 = O(\gamma^2(n)/\epsilon^2(n)). \tag{12}$$

Hence,

$$\begin{aligned}u^2(n+1) &= \|\mathbf{z}(n\tau + \tau) - \mathbf{x}^*\|^2 = \|\mathbf{z}(n\tau) - \mathbf{x}^* + \mathbf{w}(n)\|^2\\ &= u^2(n) + 2(\mathbf{z}(n\tau) - \mathbf{x}^*)'\mathbf{w}(n) + \|\mathbf{w}(n)\|^2\\ &= u^2(n) + 2\gamma(n)(\mathbf{z}(n\tau) - \mathbf{x}^*)'\nabla f(\mathbf{z}(n\tau)) + 2(\gamma(n)/\epsilon(n))(\mathbf{z}(n\tau) - \mathbf{x}^*)'\mu(n)\\ &\quad + 2(\mathbf{z}(n\tau) - \mathbf{x}^*)'O(\rho(n)) + O(\gamma^2(n)/\epsilon^2(n)).\end{aligned}$$

Now,

$$(\mathbf{z}(n\tau) - \mathbf{x}^*)'\nabla(f(\mathbf{z}(n\tau)) \le f(\mathbf{z}(n\tau)) - f(\mathbf{x}^*), \tag{13}$$

by concavity. (See Lemma 6.) Thus,

$$u^2(n+1) \le u^2(n) + 2\gamma(n)(f(\mathbf{z}(n\tau)) - f(\mathbf{x}^*)) + 2(\gamma(n)/\epsilon(n))(\mathbf{z}(n\tau) - \mathbf{x}^*)'\mu(n) + 2(\mathbf{z}(n\tau) - \mathbf{x}^*)'O(\rho(n)) + O(\gamma^2(n)/\epsilon^2(n)). \tag{14}$$

When $u(n) > \delta$, one has

$$f(\mathbf{z}(n\tau)) < f(\mathbf{x}^*) - \beta u(n) \tag{15}$$

for some $\beta > 0$, by the strict concavity of $f(\cdot)$ around $\mathbf{x}^*$. (See Lemma 7.) Also,

$$(\mathbf{z}(n\tau) - \mathbf{x}^*)'\mu(n) = u(n)\sum_k h_k(n)\mu_k(n)$$

with

$$h_k(n) = \frac{z_k(n\tau) - x_k^*}{u(n)}.$$

Hence, when $u(n) > \delta$,

$$\begin{aligned} u^2(n+1) &\le u^2(n) - 2\gamma(n)\beta u(n) + 2u(n)(\gamma(n)/\epsilon(n))\sum_k h_k(n)\mu_k(n) \\ &\quad + 2u(n)O(\rho(n)) + O(\gamma^2(n)/\epsilon^2(n)) \\ &\le u^2(n) - 2\gamma(n)\beta u(n) + 2u(n)(\gamma(n)/\epsilon(n))\sum_k h_k(n)\mu_k(n) \\ &\quad + 2u(n)O(\rho(n)) + 2u(n)O(\gamma^2(n)/\epsilon^2(n))/(2\delta) \\ &= u^2(n) - 2\gamma(n)\beta u(n) + 2u(n)(\gamma(n)/\epsilon(n))\sum_k h_k(n)\mu_k(n) \\ &\quad + 2u(n)[O(\rho(n)) + O(\gamma^2(n)/\epsilon^2(n))/(2\delta)] \\ &\le u^2(n) - 2\gamma(n)\beta u(n) + 2u(n)[(\gamma(n)/\epsilon(n))\sum_k h_k(n)\mu_k(n) + O(\kappa(n))], \end{aligned}$$

where $\kappa(n) = \max\{\gamma^2(n)/\epsilon^2(n), \rho(n))\} = \max\{\gamma^2(n)/\epsilon^2(n), \gamma(n)\epsilon^2(n)\}$.

Hence,

$$u^2(n+1) \le u^2(n) - 2u(n)[\beta\gamma(n) - \alpha(n)], \tag{16}$$

where

$$\alpha(n) = (\gamma(n)/\epsilon(n))\sum_k h_k(n)\mu_k(n) + O(\kappa(n)).$$

Now, (16) implies (9), that is,

$$u(n+1) \le u(n) - [\beta\gamma(n) - \alpha(n)].$$

Indeed, if this last inequality were violated, one would have

$$\begin{aligned} u^2(n+1) &> \{u(n) - [\beta\gamma(n) - \alpha(n)]\}^2 = u^2(n) - 2u(n)[\beta\gamma(n) - \alpha(n)] + [\beta\gamma(n) - \alpha(n)]^2 \\ &\ge u^2(n) - 2u(n)[\beta\gamma(n) - \alpha(n)]. \end{aligned}$$

This would contradict (16).

To show that $\sum\alpha(n)$ converges to a finite random variable in Lemma 8, one uses the martingale convergence theorem for the first term and the fact that $\sum_n \kappa(n) < \infty$ by (6) and (7). For the first term, the key observation is that $h_k^2(n) \le 1$. (See Lemma 8.)

When $u(n) \le \delta$, (14)

$$\begin{aligned} u^2(n+1) &\le u^2(n) + 2(\gamma(n)/\epsilon(n))(\mathbf{z}(n\tau) - \mathbf{x}^*)'\mu(n) + O(\kappa(n)) \\ &= u^2(n) + c(n) \end{aligned}$$

with

$$c(n) = 2(\gamma(n)/\epsilon(n))(\mathbf{z}(n\tau) - \mathbf{x}^*)'\mu(n) + O(\kappa(n)).$$

The martingale convergence theorem implies that the first term goes to zero, because $\|\mathbf{z}(n\tau) - \mathbf{x}^*\| \le \delta$ and $\sum\gamma^2(n)/\epsilon^2(n) < \infty$. The last term also goes to zero. (See Lemma 9.)

The next section develops some estimates.

5.2. Preliminary Calculations

We recall the following notation that avoids having to keep track of explicit constants.

Definition 1. Let $\{h(n), n \geq 0\}$ be a sequence of positive numbers. By definition, $\{O(h(n)), n \geq 0\}$ designates a sequence of random variables such that for every sample path there is some constant C such that

$$|O(h(n))| \leq Ch(n), \ \forall n.$$

The same notation is used when the variables $O(h(n))$ are deterministic and in the vector case when the inequality holds component-wise.

This definition leads immediately to the following properties. (The last one assumes $A > 0, B > 0$ and uses $O(\gamma(n)/\epsilon(n)) \leq O(\epsilon(n))$ because $\gamma(n)/\epsilon^2(n)$ is bounded, by (6).)

Lemma 1. *One has*

$$\left[O(h(n))\right]^\alpha = O(h(n)^\alpha), \ \forall \alpha > 0, \tag{17}$$

$$O(h_1(n)) \times O(h_2(n)) = O(h_1(n)h_2(n)), \tag{18}$$

$$O(h_1(n)) + O(h_2(n)) = O(\max\{h_1(n), h_2(n)\}), \tag{19}$$

$$\text{If } h_1(n) \leq C_1 h_2(n) \leq C_2 h_1(n), n \geq n_0, \text{ then } O(h_1(n)) = O(h_2(n)), \tag{20}$$

$$\max\{\epsilon(n), \gamma(n)(A + B/\epsilon(n))\} = O(\epsilon(n)). \tag{21}$$

Lemma 2. *Let $m = T_k(n)$. We claim that*

$$f(\mathbf{x}(m+1)) - f(\mathbf{x}(m)) = a_k(n)\epsilon(n)f_k(\mathbf{z}(n\tau)) + V\epsilon(n) + a_k(n)U\epsilon^2(n) + a_k(n)V'\gamma(n) + O(\epsilon^3(n)), \tag{22}$$

where U, V, V' are bounded and independent of $a_k(n)$ and $\mathcal{F}_{n-1}$ and U is zero-mean. Also, $f_k(\mathbf{z}(n\tau))$ is the partial derivative of $f(\cdot)$ with respect to x_k evaluated at $\mathbf{z}(n\tau)$.

Proof of Lemma 2. Let $m = T_k(n)$. Recall that $T_l(n) + 1$ is the experiment time of agent l during $\{n\tau, \ldots, n\tau + \tau - 1\}$, so that $T_l(n) + 2$ is its update time. The update Equations (1) and (2) imply that, for $m \in \{n\tau, \ldots, n\tau + \tau - 1\}$,

$$q_l := x_l(m) - z_l(n\tau) = \begin{cases} a_l(n)\epsilon(n), & \text{if } T_l(n) = T_k(n) - 1 \\ g_l(n)\gamma(n), & \text{if } T_l(n) \leq T_k(n) - 2 \\ 0, & \text{otherwise} \end{cases}$$

and

$$r_l := x_l(m+1) - z_l(n\tau) = \begin{cases} a_l(n)\epsilon(n), & \text{if } T_l(n) = T_k(n) \\ g_l(n)\gamma(n), & \text{if } T_l(n) \leq T_k(n) - 1 \\ 0, & \text{otherwise.} \end{cases}$$

An important observation is that the gradient estimates $g_l(n)$ for $l \neq k$ are only affected by $a_k(n)$ at and after time $m + 1$ and then used to update x_l at or after time $m + 2$. Thus, the random variables $q_l(n)$ and $r_l(n)$ for $l \neq k$ that enter in the calculations of $\mathbf{x}(m)$ and $\mathbf{x}(m+1)$ are independent of $a_k(n)$. Moreover, $a_k(n)$ is independent of $\mathbf{z}(n\tau)$.

Definition (3) shows that, for all $l = 1, \ldots, K$,

$$|g_l(n)| \leq L + G/\epsilon(n) = O(1/\epsilon(n)),$$

where L is the Lipschitz constant and G is the bound on $\eta_k(n)$.

The identities above show that $\|\mathbf{r}\| = O(\epsilon(n))$ and $\|\mathbf{q}\| = O(\epsilon(n))$ because $g_l(n)\gamma(n) = O(\epsilon(n))$ by (21). Taylor's theorem implies the following identity:

$$f(\mathbf{x}(m+1)) - f(\mathbf{z}(n\tau)) = f(\mathbf{z}(n\tau) + \mathbf{r}) - f(\mathbf{z}(n\tau)) = \mathbf{r}'\nabla f(\mathbf{z}(n\tau)) + \frac{1}{2}\mathbf{r}'H\mathbf{r} + O(\epsilon^3(n)),$$

where $H = Hf(\mathbf{z}(n\tau))$ is the Hessian of $f(\cdot)$ evaluated at $\mathbf{z}(n\tau)$.

Similarly,

$$f(\mathbf{x}(m)) - f(\mathbf{z}(n\tau)) = f(\mathbf{z}(n\tau) + \mathbf{q}) - f(\mathbf{z}(n\tau)) = \mathbf{q}'\nabla f(\mathbf{z}(n\tau)) + \frac{1}{2}\mathbf{q}'H\mathbf{q} + O(\epsilon^3(n)).$$

Subtracting these two expressions, we find

$$f(\mathbf{x}(m+1)) - f(\mathbf{x}(m)) = (\mathbf{r}-\mathbf{q})'\nabla f(\mathbf{x}(n\tau)) + \frac{1}{2}(\mathbf{r}-\mathbf{q})'H(\mathbf{r}+\mathbf{q}) + O(\epsilon^3(n)).$$

Now,

$$r_l - q_l = \begin{cases} a_l(n)\epsilon(n), & \text{if } T_l(n) = T_k(n) \\ g_l(n)\gamma(n) - a_l(n)\epsilon(n), & \text{if } T_l(n) = T_k(n)-1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$r_l + q_l = \begin{cases} a_l(n)\epsilon(n), & \text{if } T_l(n) = T_k(n) \\ g_l(n)\gamma(n) + a_l(n)\epsilon(n), & \text{if } T_l(n) = T_k(n)-1 \\ 2g_l(n)\gamma(n), & \text{if } T_l(n) \le T_k(n)-2 \\ 0, & \text{otherwise.} \end{cases}$$

In the rest of this section, U, U_1, U_2, U_3 designate random variables that are bounded, zero-mean, and independent of $a_k(n)$ and $V, V', V_1, V_2, V_3, V_4$ designate random variables that are bounded and independent of $a_k(n)$.

By examining the terms in $\mathbf{r}-\mathbf{q}$, we find that

$$(\mathbf{r}-\mathbf{q})'\nabla f(\mathbf{x}(n\tau)) = a_k(n)\epsilon(n)f_k(\mathbf{z}(n\tau)) + W,$$

where W is a sum of terms of the forms

$$a_l(n)\epsilon(n)f_l(\mathbf{z}(n\tau)) \text{ and } g_l(n)\gamma(n)f_l(\mathbf{z}(n\tau)).$$

Thus, the terms of the above two types are either of the form

$$U_1\epsilon(n) \text{ or } V_1\gamma(n)/\epsilon(n).$$

We conclude that

$$(\mathbf{r}-\mathbf{q})'\nabla f(\mathbf{x}(n\tau)) = a_k(n)\epsilon(n)f_k(\mathbf{z}(n\tau)) + U_1\epsilon(n) + V_1\gamma(n)/\epsilon(n).$$

The sum $(\mathbf{r}-\mathbf{q})'H(\mathbf{r}+\mathbf{q})$ is composed of terms that are multiples of one of the following three expressions:

$$a_i(n)a_j(n)H_{i,j}\epsilon^2(n), a_i(n)g_j(n)H_{i,j}\epsilon(n)\gamma(n), g_i(n)g_j(n)H_{i,j}\gamma^2(n).$$

Terms of first type yield a sum $a_k(n)U_2\epsilon^2(n) + V_2\epsilon^2(n)$, where $a_k(n)U_2\epsilon^2(n) = 2\sum_{j\neq k}a_k(n)a_j(n)H_{k,j}$ and $V_2\epsilon^2(n) = a_k(n)a_k(n)H_{k,k}\epsilon^2(n) = H_{k,k}\epsilon^2(n)$.

Terms of the second or third type yield a sum $U_3\gamma(n) + a_k(n)V_3\gamma(n) + V_4\gamma^2(n)/\epsilon^2(n)$.

Combining the observations above, we conclude that

$$\begin{aligned} &f(\mathbf{x}(m+1)) - f(\mathbf{x}(m)) \\ &= a_k(n)\epsilon(n)f_k(\mathbf{z}(n\tau)) + U_1\epsilon(n) + V_1\gamma(n)/\epsilon(n) + a_k(n)U_2\epsilon^2(n) + V_2\epsilon^2(n) \\ &\quad + U_3\gamma(n) + a_k(n)V_3\gamma(n) + V_4\gamma^2(n)/\epsilon^2(n) + O(\epsilon^3(n)) \\ &= a_k(n)\epsilon(n)f_k(\mathbf{z}(n\tau)) + V\epsilon(n) + a_k(n)U\epsilon^2(n) + a_k(n)V'\gamma(n) + O(\epsilon^3(n)), \end{aligned}$$

where U, V, V' are defined as

$$\begin{aligned} V\epsilon(n) &= U_1\epsilon(n) + V_1\gamma(n)/\epsilon(n) + V_2\epsilon^2(n) + U_3\gamma(n) + V_4\gamma^2(n)/\epsilon^2(n) \\ a_k(n)U\epsilon^2(n) &= a_k(n)U_2\epsilon^2(n) \\ a_k(n)V'\gamma(n) &= a_k(n)V_3\gamma(n). \end{aligned}$$

This is (22).

You will note that in this derivation, all the terms involving $g_l(n)$ are due to the asynchronous updates where some agents update while others are estimating the partial derivatives.

Lemma 3. *Let $m = T_k(n)$. We claim that*

$$g_k(n) = f_k(\mathbf{x}(n\tau)) + \mu_k(n)/\epsilon(n) + O(\gamma(n)/\epsilon(n)) + O(\epsilon^2(n)), \tag{23}$$

where $\mu_k(n)$ is a bounded random variable that is zero-mean given $\mathcal{F}_{n-1}$.

Proof of Lemma 3. Because $a_k(n) = 1/a_k(n)$ (because $a_k(n) \in \{-1, 1\}$) one has, using Lemma 2,

$$\begin{aligned} g_k(n) &= \frac{f(\mathbf{x}(m+1)) - f(\mathbf{x}(m)) + \eta_k(n)}{a_k(n)\epsilon(n)} \\ &= a_k(n)\frac{a_k(n)\epsilon(n)f_k(\mathbf{z}(n\tau)) + V\epsilon(n) + a_k(n)U\epsilon^2(n) + a_k(n)V'\gamma(n) + O(\epsilon^3(n)) + \eta_k(n)}{\epsilon(n)} \\ &= f_k(\mathbf{z}(n\tau)) + a_k(n)V + U\epsilon(n) + V'\gamma(n)/\epsilon(n) + a_k(n)\eta_k(n)/\epsilon(n) + O(\epsilon^2(n)). \end{aligned}$$

This expression is of the form (23), with

$$\mu_k(n)/\epsilon(n) = a_k(n)V + U\epsilon(n) + a_k(n)\eta_k(n)/\epsilon(n)$$

and

$$O(\gamma(n)/\epsilon(n)) + O(\epsilon^2(n)) = V'\gamma(n)/\epsilon(n) + a_k(n)O(\epsilon^2(n)).$$

5.3. Proofs of the Main Steps

The following lemma shows that (11) holds.

Lemma 4. *Let* $\mathbf{w}(n) = \mathbf{z}(n\tau + \tau) - \mathbf{z}(n\tau)$. *One has*

$$\mathbf{w}(n) = \gamma(n)\nabla f(\mathbf{z}(n\tau)) + (\gamma(n)/\epsilon(n))\mu(n) + O(\rho(n)), \tag{24}$$

where $\mu(n)$ *is a bounded random vector that is zero-mean given* $\mathcal{F}_{n-1}$ *and* $\rho(n) = \max\{\gamma^2(n)/\epsilon(n), \gamma(n)\epsilon^2(n)\}$.

Proof of Lemma 4. One has

$$z_k(n\tau + \tau) = z_k(n\tau) + \gamma(n)g_k(n), \tag{25}$$

so that Lemma 3 implies that

$$\begin{aligned} \mathbf{w}(n) &= \gamma(n)\nabla f(\mathbf{z}(n\tau)) + (\gamma(n)/\epsilon(n))\mu(n) + \gamma(n)[O(\gamma(n)/\epsilon(n)) + O(\epsilon^2(n))] \\ &= \gamma(n)\nabla f(\mathbf{z}(n\tau)) + (\gamma(n)/\epsilon(n))\mu(n) + O(\rho(n)). \end{aligned}$$

Hence, (24) holds.

The following lemma shows that (12) holds.

Lemma 5. *Let* $\mathbf{w}(n) = \mathbf{z}(n\tau + \tau) - \mathbf{z}(n\tau)$. *One has*

$$\|\mathbf{w}(n)\|^2 = O(\gamma^2(n)/\epsilon^2(n)). \tag{26}$$

Proof of Lemma 5. In (11), which is also (24), the gradient of $f(\cdot)$ is bounded and so is the vector $\mu(n)$. Hence,

$$\mathbf{w}(n) = O(\gamma(n)) + O(\gamma(n)/\epsilon(n)) + O(\rho(n)) = O(\gamma(n)/\epsilon(n)).$$

Thus, (26) holds.

The following lemma proves (13)

Lemma 6. *One has*

$$(\mathbf{z}(n\tau) - \mathbf{x}^*)'\nabla(f(\mathbf{z}(n\tau)) \le f(\mathbf{z}(n\tau)) - f(\mathbf{x}^*). \tag{27}$$

Proof of Lemma 6. A smooth concave function lies below its supporting hyperplanes. □

The following lemma proves (15).

Lemma 7. *For any* $\delta > 0$, *there is some* $\beta > 0$ *such that*

$$f(\mathbf{z}) \le f(\mathbf{x}^*) - \beta\|\mathbf{z} - \mathbf{x}^*\|, \text{ if } \|\mathbf{z} - \mathbf{x}^*\| \ge \delta. \tag{28}$$

Proof of Lemma 7. By continuity and strict concavity in a neighborhood of $\mathbf{x}^*$,

$$-\alpha := \max\{f(\mathbf{z}) - f(\mathbf{x}^*)|\, \|\mathbf{z} - \mathbf{x}^*\| \ge \delta\} < 0.$$

Let $\beta = \alpha/\delta$. Assume $\|\mathbf{z} - \mathbf{x}^*\| \ge \delta$. Define $\mathbf{v}$ as follows:

$$\mathbf{v} = \rho\mathbf{z} + (1-\rho)\mathbf{x}^* \text{ with } 1 - \rho = \delta/\|\mathbf{z} - \mathbf{x}^*\|.$$

Then,

$$\|\mathbf{v}-\mathbf{x}^*\| = \|(1-\rho)\mathbf{z}-(1-\rho)\mathbf{x}^*\| = (1-\rho)\,\|\mathbf{z}-\mathbf{x}^*\| = \delta.$$

Consequently,

$$f(\mathbf{v}) - f(\mathbf{x}^*) \le -\alpha.$$

Also, by concavity,

$$f(\mathbf{v}) \ge \rho f(\mathbf{x}^*) + (1-\rho) f(\mathbf{z}).$$

Hence,

$$\theta f(\mathbf{x}^*) + (1-\rho) f(\mathbf{z}) \le f(\mathbf{x}^*) - \alpha,$$

so that

$$f(\mathbf{z}) \le f(\mathbf{x}^*) - \frac{\alpha}{1-\rho} = f(\mathbf{x}^*) - \beta\,\|\mathbf{z}-\mathbf{x}^*\|,$$

as claimed.

The following lemma shows that the sequence $a(n)$ in (9) sums to a finite random variable.

Lemma 8. *Let*

$$\alpha(n) = (\gamma(n)/\epsilon(n))\sum_k h_k(n)\mu_k(n) + O(\kappa(n)),$$

where $\kappa(n) = \max\{\gamma^2(n)/\epsilon^2(n), \gamma(n)\epsilon^2(n)\}$.

Then the sum of $\alpha(n)$ *converges to a finite random variable.*

Proof of Lemma 8. First consider

$$(\gamma(n)/\epsilon(n))h_k(n)\eta_k(n).$$

Recall that $|h_k(n)| \le 1$ and that the random variables $\eta_k(n)$ are bounded and zero-mean given $\mathcal{F}_{n-1}$. Thus, the sum

$$\sum_{n=0}^{m}(\gamma(n)/\epsilon(n))h_k(n)\mu_k(n)$$

is a martingale with respect to that filtration $\mathcal{F}_m$. Moreover,

$$\sum_n (\gamma(n)/\epsilon(n))^2 < \infty$$

by assumption. Consequently, by the martingale convergence theorem, this sum converges to a finite random variable.

Also, the terms $O(\kappa(n))$ sum to finite numbers, by (6).

The following lemma shows that the $c(n)$ in (10) converge to zero.

Lemma 9. *Let*

$$c(n) = 2(\gamma(n)/\epsilon(n))(\mathbf{z}(n\tau)-\mathbf{x}^*)'\mu(n) + O(\kappa(n))$$

for n *such that* $\|\mathbf{z}(n\tau)-\mathbf{x}^*\| \le \delta$ *and* $c(n) = 0$ *otherwise. Then* $c(n) \to 0$.

Proof of Lemma 9. Consider the term

$$(\gamma(n)/\epsilon(n))(z_k(n\tau)-x_k^*)\mu_k(n).$$

Note that

$$\begin{aligned}|(\gamma(n)/\epsilon(n))(z_k(n\tau)-x_k^*)|^2 &\le (\gamma^2(n)/\epsilon^2(n))|z_k(n\tau)-x_k^*|^2\\ &\le (\gamma^2(n)/\epsilon^2(n))\,\|\mathbf{z}(n\tau)-\mathbf{x}^*\|^2 \le (\gamma^2(n)/\epsilon^2(n))\delta^2.\end{aligned}$$

Consequently, as in Lemma 8, these terms sum to a finite random variable. Hence, the terms converge to zero.

The terms $O(\kappa(n))$ also converge to zero.

6. Conclusions

This paper proves the convergence of a distributed version of the Kiefer-Wolfowitz algorithm under some strong assumptions. The function is assumed to be strictly concave in a neighborhood of its maximizer and with bounded derivatives up to the third order. The observation noise is assumed to be bounded. The agents update regularly, once in every interval of τ successive steps. The proof is self-contained and does not require assuming that the variables visit a compact set infinitely often. Instead, it shows that the updates prevent the variables from drifting away.

Many of these assumptions are probably stronger than necessary. For instance, the rate of updates of the different agents could be different. Convergence in probability should occur if only moments of the noise are bounded. Also, a variation where the agents estimate the function by averaging their observations between successive updates instead of in one step may be more relevant for some applications and should be explored. The situation where the agents choose to experiment at random times is also important for applications but seems to require a different proof technique. Finally, a projection version of the algorithm and the case of discrete variables are left for further study.

Acknowledgments

This work was motivated by an application to wireless networks studied with Piotr Gawlowicz and Adam Wolisz. They identified the importance of asynchronous distributed updates in that application. The author is grateful for their suggestions for this paper and to the associate editor and the referees for their judicious, constructive, and prompt comments.

References

Kennedy RKL, Khoshgoftaar TM, Villanustre F, Humphrey T (2019) A parallel and distributed stochastic gradient descent implementation using commodity clusters. *J. Big Data* 6(1):1–23.

Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* 23(3):462–466.

Kushner H, Clark DJ (1978) *Stochastic Approximation Methods for Constrained and Unconstrained Systems* (Springer, New York).

Ljung L, Söderström T (1983) *Theory and Practice of Recursive Identification* (MIT Press, Cambridge, MA).

Nedic A, Ozdaglar A (2009) Distributed subgradient methods for multi-agent optimization. *IEEE Trans. Automatic Control* 54(1):48–61.

Ramaswamy A (2019) DSPG: Decentralized simultaneous perturbations gradient descent scheme. Preprint, submitted March 17, https://arxiv.org/abs/1903.07050.

Spall J (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automatic Control* 37(3):332–341.

Swenson B, Murray R, Kar S, Poor V (2020) Distributed stochastic gradient descent: Nonconvexity, nonsmoothness, and convergence to local minima. Preprint, submitted March 5, https://arxiv.org/abs/2003.02818.