Distributable Computations

Jeffrey A. Barnett

August 26, 2015

1 Summary

There are two rather different motivations to distribute a computation. The first is to reduce the time to obtain answers by applying multiple computers in parallel. In recent times, for example, computer owners all over the world have donated resources to the search for large prime numbers. The use of thousands of computers has reduced the time between successes to months rather than years. The second motivation for distribution is to use computation to reduce the bandwidth necessary to transport data. In some cases, multiple local sites contribute to a central site where a fused result is produced. Summary statistics and reductions of large data collections are typical calculations of this sort. The computations are not particularly taxing so communications resources represent the major cost factor. Only the second motivation is considered herein.

A conjecture about necessary and sufficient conditions that the computation of a function can be distributed to reduce bandwidth is stated below. That conjecture concerns equivalence classes—partitions—of a function's domain. The elements, x_1 and x_2, are in the same partition if $f(x_1, y) = f(x_2, y)$ for all y, where x_1, x_2, and y are tuples of data. The conjecture is formed in terms of the "dimensionality" of a surface that intersects each member of the partition.

The next section presents a simple example, the computation of the mean and variance of a sample split between two sites, to motivate the formal definition of distributable computation introduced in Section 3. The following section makes a conjecture, that if true, would provide an alternative characterization of distributable functions. Section 5 then casts the conjecture in terms of partitions of the function's domain and a set of representatives for

elements of the partition. In addition, it is noted why proving the conjecture might be difficult. Finally, Section 6 examines the class of monotonic functions where the conjecture is true and, therefore, a straightforward method to check whether a computation is distributable or not is available.

2 An Example

An example is used to motivate the definition of distributed computation introduced below. The computation to be distributed is (A, V), where A is the average and V is the variance of the numbers $s_1, \ldots, s_n$.

$$A = \sum_{i=1}^{n} s_i/n \qquad\qquad V = \sum_{i=1}^{n}(s_i - A)^2/n.$$

These results can easily be computed from

$$\alpha = \sum_{i=1}^{n} s_i \qquad\qquad \beta = \sum_{i=1}^{n} s_i^2$$

by the formulas $A = \alpha/n$ and $V = \beta/n - A^2$. Assume that the numbers $s_1, \ldots, s_k$ are resident at one site while $s_{k+1}, \ldots, s_n$ are resident at another. One possibility is to transmit the k-tuple, $(s_1, \ldots, s_k)$, to the second site where A and V will be computed. However, there is a more economical possibility shown in Figure 1. Calculate α_k and β_k at the first site (CPU_1)

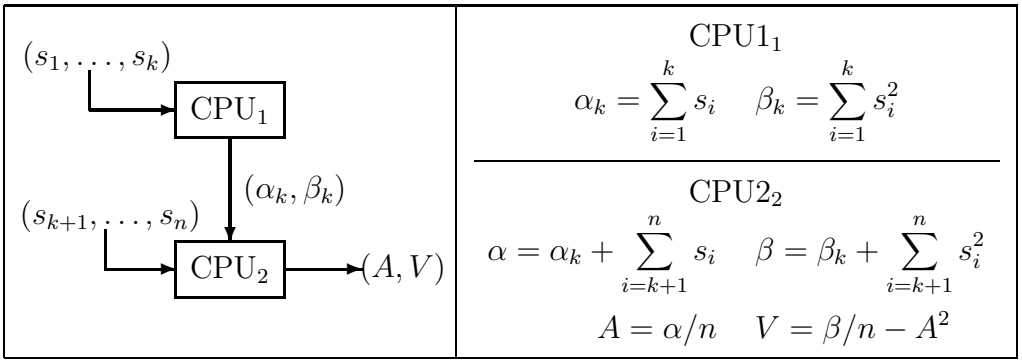


Figure 1: Distributed computation of average and variance of $s_1, \ldots, s_n$.

and only transmit these two quantities to the second site (CPU_2) where α and β, then A and V are computed. In this example, the bandwidth is reduced by a factor of $1 - 2/k$ which can be considerable if k is large. Achieving this sort of reduction is the motivation for the distributed computation model described next.

3 Model of Distributed Computation

The following definition, as depicted in Figure 2, is meant to capture the idea of distributing a computation to reduce bandwidth as discussed above.

Definition 1 *$f\colon X \times Y \to T$ is* Z-distributable in X *if there is a continuous onto $d\colon X \to Z$ and a continuous $c\colon Z \times Y \to T$, such that $f(x, y) = c(d(x), y)$ for all $x \in X$ and $y \in Y$.*

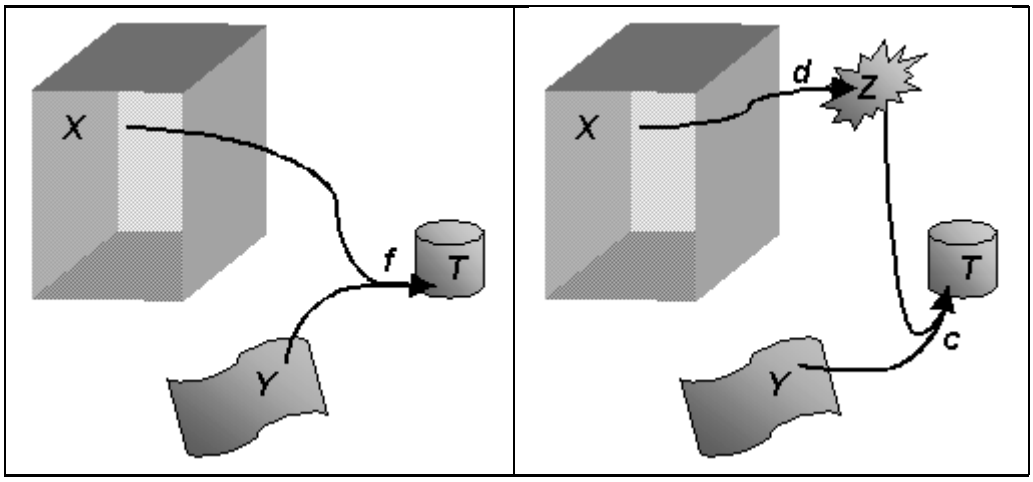


Figure 2: Computation $f(x, y)$ is Z-distributable as $c(d(x), y)$.

The definition entails that f be continuous because it is a composition of continuous functions and that Z be connected when X is connected because d is continuous. (Connectivity of X is assumed below.) Continuity of functions is required for two reasons. The first is that *computation* of discontinuous functions is rare and somewhat ill-defined because computation entails truncation and, hence, approximations. The second is that *communications* of data by a discontinuous function—say by a reduction operation from $\Re^n$ to $\Re$ that interlaces decimal digits—is not germane to actual computations and finite bandwidth resources.

The cases of interest are those where $X \subset \Re^n$, $Z \subset \Re^u$, and $u < n$. Of course there is always a Z such that f is Z-distributable if $u = n$. A sharper definition is suggested though it isn't used below.

Definition 2 *The function $f\colon X \times Y \to T$ is* exactly Z-distributable in X, *where $Z \subset \Re^u$, if it is Z-distributable and there is no $v < u$ and $Z' \subset \Re^v$ such that it is Z'-distributable.*

Consider the example from the previous section in terms of the first definition: Let $X = \Re^k$, $Y = \Re^{n-k}$, $T = \Re \times \Re^+$, where $\Re^+$ is the nonnegative reals, and $Z = \Re \times \Re^+$; then $d(s_1, \ldots, s_k) \rightarrow (\alpha_k, \beta_k)$ and $c((\alpha_k, \beta_k), s_{k+1}, \ldots, s_n)$ are the computations shown in Figure 1.

The median is an example of a function that is difficult to distribute to advantage. It is defined as $f(x_1, \ldots, x_{2n+1}) = x_{j_{n+1}}$, where $j_1, \ldots, j_{2n+1}$ is a permutation of $1, \ldots, 2n+1$, such that $x_{j_i} \leq x_{j_{i+1}}$. Consider the potential distributed computation,

$$f(x_1, \ldots, x_{2n+1}) \;=\; c(d(x_1, \ldots, x_{n+1}), x_{n+2}, \ldots, x_{2n+1}),$$

where $d\colon \Re^{n+1} \rightarrow \Re^v$. The questions is, how small can v be? The answer is that $v \geq n+1$ because the $x_{n+2}, \ldots, x_{2n+1}$ can be chosen such that *any one* of $x_1, \ldots, x_{n+1}$ is the value of f. Since no continuous d can exactly encode independent $x_1, \ldots, x_{n+1}$ in less than $n+1$ real numbers, the answer follows.

Now assume that $2n+1$ numbers are split between two sites, $0 \leq m \leq 2n+1$ at one site and $2n+1-m$ at the other, and it is desired to compute the median using the least possible communications. It can be shown, using an argument similar to the above, that at least $b = \min(m, 2n+1-m)$ numbers must be transmitted. The minimum is achieved by sending all b of the numbers from the site with the smallest collection to the other site where the median will be computed. If the median is to be calculated at the smaller site, $b+1$ numbers must be transmitted.

4 A Conjecture

A conjecture about Z-distributable functions is formed in terms of a partition induced on X by f. Assume $f\colon X \times Y \rightarrow T$ as above, and define

$$\begin{aligned} F_x &= \{q \in X \mid \forall y \in Y : \{f(q,y) = f(x,y)\}\} \\ \mathcal{F} &= \{F_x \mid x \in X\}. \end{aligned}$$

Thus, $\mathcal{F}$ is a partition of X and the elements of an $F_x \in \mathcal{F}$ are indistinguishable from x vis-á-vis f.

Theorem 3 *If $f\colon X \times Y \rightarrow T$, $W \subset X$, $r\colon X \rightarrow W$ is continuous onto, where $r(x) \in F_x$, and $p\colon W \rightarrow Z$ is continuous 1-to-1, then f is Z-distributable.*

Proof Let $d(x) = p(r(x))$ and $c(z,y) = f(p^{-1}(z), y)$. Then $f(x,y)) = c(d(x), y)$ because the conditions guarantee that p^{-1} is well defined. QED

Conjecture 4 *If f is Z-distributable, then there is a $W \subset X$, a continuous onto $r\colon X \to W$, where $r(x) \in F_x$, and a continuous 1-to-1 $p\colon W \to Z$.*

The conjecture if true, together with the theorem, would provide an alternative, equivalent definition of Z-distributable functions.

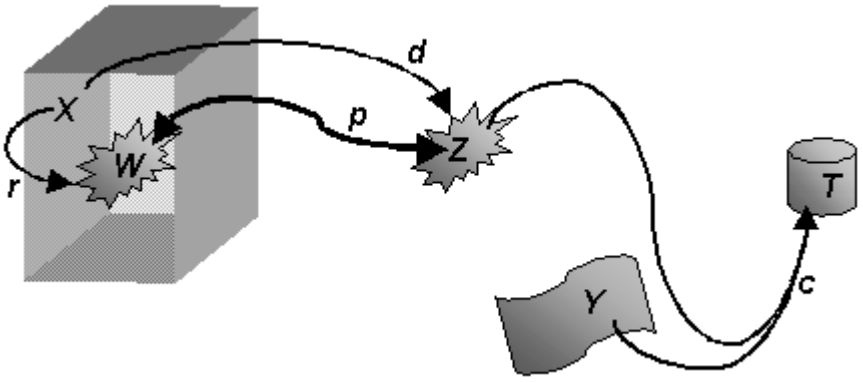


Figure 3: The conjecture is that Z-distribution by $c(d(x), y)$ entails a homomorphism p of Z and W, and $r\colon X \to W$ where $r(x) \in F_x$.

It is straightforward to find W, r, and p that satisfy the conjecture for the computation of the mean and variance as described in Sections 2 and 3. Let $s = (s_1, \ldots, s_k)$ and $t = (t_1, \ldots, t_k)$. Then, using the formulas shown in Figure 1, $t \in F_s$ if and only if $\sum t_i = \sum s_i$ and $\sum t_i^2 = \sum s_i^2$. Let

$$W = \{(a - \delta, a, \ldots, a, a + \delta) \mid a \in \Re \wedge \delta \in \Re^+\}$$

and define $p\colon W \to Z$ as $p(w) = p(a - \delta, a, \ldots, a, a + \delta) \to (a, \delta)$. Clearly, p is continuous and 1-to-1 onto $Z = \Re \times \Re^+$ as required. It remains to find a continuous $r\colon X \to W$, onto, where $r(x) \in F_x$. Simply define $r(x_1, \ldots, x_k) \to (a - \delta, a, \ldots, a, a + \delta)$, where $a = \alpha/k$, $\delta = \sqrt{(k\beta - \alpha^2)/(2k)}$, $\alpha = \sum x_i$, and $\beta = \sum x_i^2$. The fact that $r(x) \in F_x$ is easily verified by substitution.

5 Discussion

In the theorem and conjecture, $W \subset X$ has a special significance—it provides a set of representatives for a partition of X that refines $\mathcal{F}$. Let

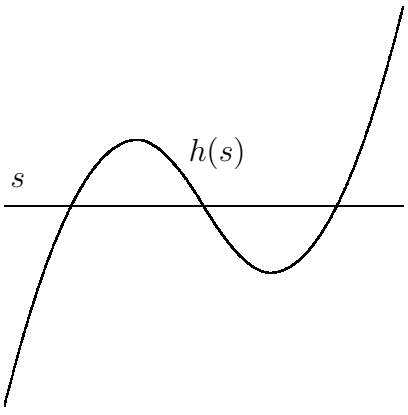
$$W_x = \{q \in X \mid r(q) = r(x)\}$$
$$\mathcal{W} = \{W_x \mid x \in X\}$$

$\mathcal{W}$ clearly is a partition of X and it refines $\mathcal{F}$ because

$$\forall x, q \in X : \{W_x \subset F_q \vee W_x \cap F_q = \emptyset\}.$$

Essentially, p^{-1} maps Z into X such that its image, W, hits each element of $\mathcal{F}$ at least once and r maps each $x \in X$ to a point in W that is also in F_x. The conjecture is that this is always possible if f is Z-distributable. In other words, if f is Z-distributable, the conjecture postulates a contraction, r, of X to a homomorphic image, W of Z, that preserves the partitioning induced on X by f, i.e., $r(x) \in F_x$.

In some cases it is possible to construct the W, r, and p, required by the conjecture from the d used to distribute f. If there is a 1-to-1 continuous $q\colon Z \to X$ such that $d(q(z)) = z$, this is certainly the case: simply define $W = q(Z)$, $p(w) = q^{-1}(w)$, and $r(x) = q(d(x))$. Unfortunately, such a q does not in general exist for a given d (see Figure 4) so this is not a fruitful



$f(x_1, x_2, y) = yh(x_1)$ is $\Re$-distributable by $f(x_1, x_2, y) = c(d(x_1, x_2), y)$ where $d(x_1, x_2) = h(x_1)$ and $c(z, y) = zy$. However, there is no 1-to-1 continuous q such that $d(q(z)) = z$ because h does not have a continuous inverse over its entire range. However, if f is distributed by $d(x_1, x_2) = x_1$ and $c(z, y) = yh(z)$, simply define $q(z) = (z, 0)$ and the conjecture is seen to hold for this case.

Figure 4: A problem with straightforward verification of the conjecture

avenue to seek a general proof of the conjecture. However for some classes of functions, it is easy to show that the required q can be constructed from any d used to distribute the computation. An example is presented next.

6 Distributing Monotonic Functions

Let $f\colon X \times Y \to \Re$ be Z-distributable, where $X = S^n$, $S \subset \Re$ is an interval, and $Z \subset \Re$. If $f(s_1, \ldots, s_n, y)$ is strictly monotonic in each s_i, the q described

in the previous section can be constructed as follows. Define W as the "main diagonal" of the hypercube, X, i.e.,

$$W \;=\; \{(s,\ldots,s) \mid s \in S\}.$$

Let d_W be d restricted to W. It will be shown that (1) d_W is 1-to-1 into Z and (2) d_W is onto Z. Hence, d_W^{-1} exists and is our q and, therefore, the conjecture will be established for this restricted class of monotonic functions. Consider

$$c(d(s,\ldots,s),y) \;=\; f(s,\ldots,s,y).$$

Since the right hand side is strictly monotonic in s, d_W must be strictly monotonic in s too and, hence, is 1-to-1 with a well-defined inverse from $d_W(W)$ to W. To show that d_W is onto Z, I will prove that for an arbitrary $x \in X$, there is a $w \in W$ such that $d(x) = d(w)$. Assume that f and d_W are increasing functions. If either is decreasing, a similar demonstration is available.

Let $x = (s_1,\ldots,s_n)$ be an arbitrary $x \in X$. If $x \in W$, there is nothing to show, so assume that $x \notin W$, i.e, not all of the s_i are equal, and define $m = \min s_i$ and $M = \max s_i$. Thus, there is at least one $s_i \neq m$ and one $s_i \neq M$ so

$$f(m,\ldots,m,y) \;<\; f(x,y) \;<\; f(M,\ldots,M,y),$$

for any $y \in Y$, because of monotonicity. If $d(m,\ldots,m) \leq d(x) \leq d(M,\ldots,M)$, the intermediate value theorem guarantees the existence of a $m \leq \beta \leq M$ such that $d(x) = d(\beta,\ldots,\beta)$ and clearly $(\beta,\ldots,\beta) \in W$.

The remaining cases are $d(x) < d(m,\ldots,m)$ and $d(M,\ldots,M) < d(x)$. Assume the latter—the demonstration for the first case is virtually identical. Define $g_i(t) = (1-t)m + ts_i$ and $g(t) = (g_1(t),\ldots,g_n(t))$. Note that $g(0) = (m,\ldots,m)$, $g(1) = x$, and g is continuous. Note also that $f(g(t),y)$ is strictly increasing in t. Since $d(g(0)) < d(M,\ldots,M) < d(x) = d(g(1))$, there must exist $0 < v < 1$ such that $d(g(v)) = d(M,\ldots,M)$ by the intermediate value theorem. This in turn implies that $f(M,\ldots,M,y) = f(g(v),y) < f(x,y)$. But this is a contradiction. Therefore, d_W is onto Z. The following theorem sums up this result.

Theorem 5 *The function $f\colon X \times Y \to \Re$, where $X = S^n$, $S \subset \Re$ is an interval, W is the main diagonal of X, and f is strongly monotonic in X, is Z-distributable, $Z \subset \Re$, if and only if there is a continuous onto $r\colon X \to W$, where $r(x) \in F_x$.*

In other words, f is Z-distributable, $Z \subset \Re$, if and only if there is a continuous $\rho = \rho(x_1, \ldots, x_n)$ such that $f(x_1, \ldots, x_n, y) = f(\rho, \ldots, \rho, y)$ for all $y \in Y$. This observation often provides a simple method to determine the distributability of a monotonic function. First consider a negative example:

$$f(x_1, x_2, y_1, y_2) \;=\; x_1 y_1 + x_2 y_2$$

defined for positive y_i. The idea is to find a $\rho = \rho(x_1, x_2)$ such that

$$f(x_1, x_2, y_1, y_2) \;=\; f(\rho, \rho, y_1, y_2).$$

But $\rho = (x_1 y_1 + x_2 y_2)/(y_1 + y_2)$ which necessarily depends on y_1 and y_2. Therefore, this function cannot be Z-distributed, $Z \subset \Re$.

The same method provides a demonstration that the Hölder means are $\Re^+$-distributable. These means are defined for each $v \in \Re$ as

$$h_v(x_1, \ldots, x_n) \;=\; \lim_{z \to v} \left(\sum_{i=1}^{n} x_i^z \;\Big/\; n \right)^{1/z},$$

where the x_i are positive. The limit operator is necessary for $v = 0$ which corresponds to the geometric mean. Consider distributing h_v where $x_1, \ldots, x_m$, $m < n$, are the remote elements. To show that h_v is $\Re^+$-distributable in these variables, it is only necessary to find a $\rho = \rho(x_1, \ldots, x_m)$ such that

$$\begin{aligned} h_v(x_1, \ldots, x_n) &= h_v(\rho, \ldots, \rho, x_{m+1}, \ldots, x_n) \\ \lim_{z \to v} \left(\sum_{i=1}^{n} x_i^z \;\Big/\; n \right)^{1/z} &= \lim_{z \to v} \left(\left(\sum_{i=1}^{m} \rho^z + \sum_{i=m+1}^{n} x_i^z \right) \;\Big/\; n \right)^{1/z}. \end{aligned}$$

Clearly there is a solution,

$$\rho \;=\; \lim_{z \to v} \left(\sum_{i=1}^{m} x_i^z \;\Big/\; m \right)^{1/z},$$

that is a continuous function of $x_1, \ldots, x_m$.