On "Simplifying Analyses of Chemical Reaction Networks for Approximate Majority", by Condon et al.

Seminar Advanced Algorithms and Data Structures

Luca Mondada

November 11, 2018

1 Introduction

Condon, Hajiaghavi, Kirkpatrick and Maňuch propose a new analysis of a well-known solution to the Approximate Majority in [1]. The Approximate Majority problem is a relaxation of the majority problem, where the correctness of the result only holds in probability and may depend on the difference between the number of elements in the two categories. This problem appears in molecular biology and is one of the simplest examples of a *population protocol*.

1.1 What is a population protocol?

Approximate majority works in a model of computation called population protocol. This model is well-known from chemistry: consider a typical reaction

$$CO_2 + H_2O \to H_2CO_3$$

A population protocol is based on a set of rules $X + Y \to Z$, similar to the one above. We call the types X, Y, Z states. From an initial mix of elements of the different states, the computation takes place by following the rules (imagine molecules that are mixed in a solution, where then reactions happen).

This model has several real world applications. The most obvious example is chemistry, but also molecular biology and in ecology to simulate population dynamics.

2 The Approximate majority problem

In this report, we will analyse one of the most simple population protocols: approximate majority. In the approximate majority problem, we use two states of elements, which we call X and Y. We will denote x, respectively y, for the number of elements of state X, respectively Y. We will further write n for the total number of elements. There are

different ways of defining population protocols to solve the approximate majority problem, with different properties and runtimes. We will first discuss the simplest version, which we call the two-state majority. We will then introduce and present the analysis by Condon et al. of the three-state majority.

2.1 The two-state majority

The two-state majority uses states X and Y and the following rules:

$$X + Y \to X + X$$
$$Y + X \to Y + Y$$

Note that in this definition the order of X and Y is important (which doesn't make sense if we think of the application in chemistry). We could also say that the order doesn't matter but the rules are non-deterministic with both rules being equally likely.

What these rules say essentially is that when an X and Y meet, one of them will transform into the other's state. If we wait for long enough, then we expect an equilibrium to form (i.e. eventually, all elements will be either all X, or all Y). We further expect that if there are more X's than Y's at the beginning, the elements will on average tend to all become X's. Thus, this solves the approximate majority problem, with the simplest rules imaginable!

This can indeed be shown to be true. However, the guarantees are quite weak: the probability of X becoming the majority if there are x elements of state X at the beginning are $\frac{x}{n}$. This is not better than picking one element at random in the mix and claiming that that state is the majority! Furthermore, the number of interactions needed with this set of rule can be shown to be $O(n^2)$. We will now see that we can do better than this naive approach.

2.2 The three-state majority

The main issue with the two-states majority described above was that the "mutation" of X into Y's (or the other way around) happened randomly: depending on the order of the elements in the reaction, the elements mutated into X's or Y's, but the number of elements of the respective state had no influence on which mutations happened.

We can improve on this by introducing a new state alongside X and Y. We introduce elements of state B. These B's can be thought of elements undecided between X and Y, called "blank". Initially, all elements are either X or Y, but adapted rules will introduce B's into the mix:

$$X + Y \to B + B$$
$$X + B \to X + X$$
$$Y + B \to Y + Y$$

Now, when an element X meets a Y, they will both be mutated into B's. These B's in turn mutate into X's and Y's. However, this time, the probability of mutating into X or Y will depend on the relative number of X's and Y's in the mix. Thus B's will be more likely to mutated into a X if X has a majority, reinforcing this majority further.

It turns out that proving an upper bound on the number of interactions needed to achieve consensus – i.e. all elements are either X or Y – is quite challenging. In a very lengthy proof, Angluin et al showed the number of interactions can be bounded with $O(n \log n)$, and that the consensus will be the correct solution with high probability provided that the initial difference between the number of elements of state X and Y, $x - y$, is in $o(\sqrt{n})$ [2]. It seemed astonishing that such a simple system requires such a lengthy proof. In their paper, Condon et al give a simpler proof for this bound, and reduce the requirement for $x - y$ to be in $\Omega(\sqrt{n \log n})$.

In the rest of this report, we will discuss the approach taken in the paper and give an outline of the proof.

Given certain interaction rules, it has be shown that such interacting systems (a) reach a concensus – that is when all objects belong to the same species – within a given number of interactions, and (b) the species of the concensus is the species that had the majority in the beginning, given that the gap between the majority and minority was wide enough. The characteristic (a) is often called *efficiency*, and characteristic (b) is called *correctness*.

While the definition of most such interaction rules is fairly straight-forward, the proof of their correctness and, in particular, of their efficiency have proven quite challenging. The paper argues it provides a simpler proof of correctness and efficiency for a particular set of interaction rules proposed by Angluin et al [2].

2.3 Outline of the work

The proof is based on the idea of starting from a slightly modified population protocol, and then reducing this to our original population protocol.

The modified population protocol is in fact a tri-molecular CRN (chemical reaction network). We will not go into the details of the differences between a population protocol and a CRN. A tri-molecular CRN is essentially a population protocol that takes three elements on the left side of the population protocol rule.

In the next section, we will start by presenting the analysis of the tri-molecular CRN instance. We will then adapt these results in section 4 for two different bi-molecular CRNs and show how these interaction rules translate to the same rules as in the population protocol. We will conclude with a discussion of the results in section 5.

3 Tri-molecular CRN for approximate majority

We will start by considering a tri-molecular CRN composed of the following two reactions:

$$X + X + Y \to X + X + X \tag{1}$$
$$X + Y + Y \to Y + Y + Y \tag{2}$$

The CRN contains the two states X and Y. Remember that we write x and y for the number of elements in the respective state. We will assume without loss of generality $x > y$, as the interaction rules are symmetric with regard to X and Y. Please note further that unless explicitly stated otherwise, all stated bounds and results hold with high probability (i.e. there is a chance of the result being incorrect, but this tends to 0 as n tends to infinity).

We want to show that given these two reactions, a mix rapidly reaches a majority concensus, given an initial gap of $x - y = \Omega(\sqrt{n \log n})$.

3.1 Phases and stages

Assuming $x > y$, we can frame our goal of reaching a consensus as bringing y to 0. In fact, as soon as $y = 0$, all elements must be of state X and we have reached a consensus. Thus, looking at what we call the *history* of y from the first interaction in the CNR to the last, we must prove that y eventually reaches 0.

The idea behind the proof presented is to separate the history of y in phases. For each phase, we will show that (a) once the history of y enters that phase, it will never (with high probability) go back to the previous phase, and (b) it will (with high probability) move towards the next phase or stage within a certain number of interactions. The last phase ends with $y = 0$, which means we have reached a consensus.

Once that is established, it is easy to assemble the different bounds to obtain a general result on the correctness and efficiency of the proposed approach.There are three phases characterised by different regimes in the ratio of x to y. For the proof, phases 1 and 2 are further sliced into $\Theta(\log n)$ *stages*. Similar to the phases, we will aim to show that within a stage, the history of y moves from one stage to the next without ever going back (with high probability). The boundaries between phases and stages are chosen in such a way that the upper bound on the number of interactions is optimal.

Note that successive phases (and stages) overlap. That is necessary because we want to make sure that the history of y does not go back to a previous phase (or stage): we need to leave some free space in case y starts to increase as soon as it entered the new phase or stage. If the phases did not overlap, then we could not exclude that as soon as y entered the new phase, it wouldn't increase immediately and need to be shifted back to the previous phase. By adding overlap, we can make sure we stay in the new phase (with high probability), even if y increases.

The regimes and boundaries of the phases and stages are summarised in the first three columns of Table 1.

Table 1: An overview of the different phases, their stages and the bounds that apply to them with high probability.

	Regime	Boundaries	Stages	Efficiency (using Chernoff)	No-return (using Lemma 1)
Phase 1	$c_\gamma/2\sqrt{n \log n}$ $< x - y \leq$ $n(d_\gamma - 2)/d_\gamma$	ends when $y \leq n/d_\gamma$	$t = \Theta(\log n)$, starts: $x \geq y + 2^t\sqrt{n \log n}$, ends: $x \geq y + 2^{t+1}\sqrt{n \log n}$	reaches next stage within λn reactions with high proba	$x - y >$ $2^{t-1}\sqrt{n \log n}$ with high proba
Phase 2	$e_\gamma \log n$ $< y <$ $2n/d_\gamma$	ends when $y \leq e_\gamma \log n$	$s = \Theta(\log n)$, starts: $y \leq n/2^s$, ends: $y \leq n/s^{s+1}$	reaches next stage within $\lambda n/2^s$ with high proba	$y < n/2^{s-1}$ with high proba
Phase 3	$0 \leq y < 2e_\gamma \log n$	ends when $y = 0$	None	reaches $y = 0$ within $\lambda \log n$ reactions with high proba	$y < 2e_\gamma \log n$ with high proba

3.2 Mathematics and statistics tools

In this section we will review some basic results in statistics that will be relevant for our proof.

We give a result on biased random walks. A random walk (in our case: one-dimensional) is a random sequence of steps going up or down with a fixed step size, starting at a given point. Taking the integers as example, one could start at 5, and then take a random sequence of either +2 steps or -2 steps, which could give the following sequence: 5, 7, 9, 7, 5, 3... A biased random walk is a random walk where the different options (the steps) have different probabilities.

Lemma 1 (One-dimensional biased random walk [3]) *If we run an arbitrarily long sequence of independent trials, each with success probability at least p, then the probability that the number of failures ever exceeds the number of successes by b is at most* $(\frac{1-p}{p})^b$

For completeness, we mention a famous result from statistics that will be very useful in our proof.

Lemma 2 (Chernoff tail bounds [4]) *If we run N independent trials, with success probability p, then S_N, the number of successes, has expected value $\mu = Np$ and, for $0 < \delta < 1$,*

a) $P[S_N \leq (1 - \delta)\mu] \leq \exp(-\frac{\delta^2\mu}{2})$,

b) $P[S_N \geq (1 + \delta)\mu] \leq \exp(-\frac{\delta^2\mu}{3})$.

3.3 Results

As mentioned above, all we need to show is that the history of y moves from one stage to the next, and from one phase to the next, without ever going back (with high probability). This breaks down into showing that in each phase or stage, (a) *no-return* holds, that is, the history of y will never move back to the previous phase or stage again, and

(b) *efficiency* is guaranteed, that is, the history of y will move to the next phase or stage within a bounded number of reactions.

All the relevant bounds are summarised in Table 1. Note that the derivation of the results also provide an expression for the "high probability" that increases with the number of molecules n. These are however not reported in the table for the sake of conciseness. We refer the interested reader to the original paper [1].

We will not prove the results for every phase and stage, as the results and proofs are very similar. We will focus on deriving the results for phase 1.

Lemma 3 *At any point in the computation, if $x - y = \Delta$ then the probability that $x - y \leq \Delta/2$ at some subsequent point in the computation is less than $(1/e)^{\Delta^2/(2n+2\Delta)}$.*

Proof. The change in $x - y$ is the result of a biased random walk, starting at Δ and with success rate $p = x/n \geq 1/2 + \Delta/(4n)$ (which corresponds to a reaction (1)). To reach $x - y = \Delta/2$, we need $\Delta/2$ more failures than successes. Using theorem 1, we find that this is reached with probability $(\frac{1}{1+\Delta/n})^{\Delta/2} \leq (1/e)^{\Delta^2/(2n+2\Delta)}$. ■

Corollary 4 (No-return in phase 1) *In stage t of phase 1, $x-y$ reduces to $2^{t-1}\sqrt{n \log n}$ with probability less than $1/n^{2^{2t-2}}$.*

This is the first result we were looking for! We see that the lemma gives an explicit probability that tends to 0 as n increases. We thus have no-return for each stage of phase 1.

We now show efficiency for phase 1.

Lemma 5 *If $x - y = \Delta < n/2$ at some point and if $x - y$ never reduces to $\Delta/2$ (no-return), the probability that $x - y$ increases to 2Δ within λn reactions is at least $1 - \exp -\frac{(\lambda-2)\Delta^2}{\lambda(2n+\Delta)}$.*

Proof. Similar to theorem 1, this time using Chernoff. The probability of completing λn reactions with fewer than $\lambda n/2 + \Delta/2$ successes (follows from $x - y < 2\Delta$) is at most $\exp -\frac{(\lambda-2)\Delta^2}{\lambda(2n+\Delta)}$ ■

Corollary 6 (efficiency in phase 1) *In stage t of phase 1, assuming that $x - y$ never reduces to $2^{t-1}\sqrt{n \log n}$, the probability that $x - y$ increases to $2^{t+1}\sqrt{n \log n}$ within at most λn reactions is at least $1 - exp(-\frac{(\lambda-2)2^{2t} \log n}{3\lambda})$.*

This is the efficiency result! Combining the two results, we can say that with high probability, the history of y will go from stage to stage until entering phase 2.

The same results with similar bounds can be obtained for phases 2 and 3.

4 Bi-molecular emulation of approximate majority

We now have bounds for a tri-molecular CRN. Our goal is to use this result for the three-state population protocol as proposed by Angluin et al [2]. We will see that the

same analysis of section 3 for the tri-molecular CRN can be applied to the population protocol described in section 2 with small adaptations.

Recall the rules of the three-state population protocol:

$$X + Y \to B + B \tag{0'}$$
$$X + B \to X + X \tag{1'}$$
$$Y + B \to Y + Y \tag{2'}$$

We will start with the same configuration as in the tri-molecular CRN, meaning that $b = 0$ in the initial configuration.

The trick to analyse this population protocol is to introduce the new variables $\hat{x} = x + b/2$ and $\hat{y} = y + b/2$. We see immediately that reaction (0') leaves these new variables unchanged. Reactions (1') and (2') on the other hand change the new variables at exaclty half the rate that their respective counterparts changed x and y in the previous tri-molecular CRN. We are thus in a very similar setting as in the tri-molecular case, where we simply need twice the number of interactions. To conclude the proof, one needs to show that the probabilities of reactions (1') and (2') happening can be bound by a constant times the equivalent reactions in the three-state population protocol. This allows to prove both correctness and efficiency of the emulation.

5 Conclusion

Condon et al. propose a new approach to derive bounds on the Population Protocol introduced by Angluin [2]. They start with a tri-molecular CRN that is easier to solve and then show that the same analysis can easily adapted to three-state population protocol version of the problem.

This approach of translating the population protocol into a CRN and then modifying the CRN into another more easily solvable CRN could have a lot of potential for other still open population protocols. Further, this reduction method could also be used to reduce open CRN problems to already known results, instead of proving new bounds every time.

References

[1] Condon, Anne, et al. "Simplifying Analyses of Chemical Reaction Networks for Approximate Majority." *International Conference on DNA-Based Computers.* Springer, Cham, 2017.

[2] Angluin, Dana, James Aspnes, and David Eisenstat. "A simple population protocol for fast robust approximate majority." *Distributed Computing* 21.2 (2008): 87-102.

[3] Bruguire, Catherine, Andre Tiberghien, and Pierre Clment, eds. *Topics and Trends in Current Science Education: 9th ESERA Conference Selected Contributions.* Vol. 1. Springer Science & Business Media, 2013.

[4] Chernoff, Herman. "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations." *The Annals of Mathematical Statistics* 23.4 (1952): 493-507.