Testing without Polya Urns

Rachael Caelie (Rocky) Aikens

12/22/2020

Introduction

In many ways, the doctors in this simulation will do something very clearly foolish: they will neglect to account for selection bias as they learn from past data. In the field of reinforcement learning, this would be very quickly identified as a bad algorithm for learning from data. In the field of medicine, this kind of mistake happens perhaps much more often than we would like to admit. When we build models to identify risk factors for a disease or the symptoms of that disease - either with machine learning or with more classical statistics - we seldom correct for the processes that direct who is *considered* for the diagnoses we use as our "ground truth."

Set-Up

Suppose that a doctor is trying to decide whom to consider for followup for a disease diagnosis. If they believe there is a high enough probability that the patient in front of them has the disease, they will follow up with a test or a referal to the specialist. We'll assume for the moment that the test or specialist is perfect - a person gets a positive test result or a positive diagnosis from a specialist if and only if they have the disease.

Suppose also that there are two different types of people seeking the test - for example men and women - but the underlying rate of disease is not known in either group. Let λ_A and λ_B represent the rates of disease in each group.

Suppose a new patient from group comes into the office asking for a test. We'll represent their group with $g \in \{A, B\}$. The doctor selects whether to test this patient based on the existing evidence as follows:

- Let n_A and n_B represent the number of observed cases in groups A and B.
- The doctor tests the patient with probability $\hat{\Lambda}_g = \frac{n_g}{n_A + n_B}$.

The reasoning for this behavior is somewhat Bayesian. Based on Bayes Law:

$$P(Disease = 1|Group = g) \propto P(Group = g|Disease = 1)P(Disease = 1)$$

.

Why this decision-making rule? Essentially, if $P(Disease = 1)$ represents the doctor's prior belief about the probability the patient has the disease (irrespective of their group), then the Bayesian doctor might adjust this probability in light of the patient's group by multiplying by $P(Group = g|Disease = 1)$. Effectively, the doctor will be more likely to believe that the patient has the disease if the patient comes from a group which is well-represented among people with the disease.

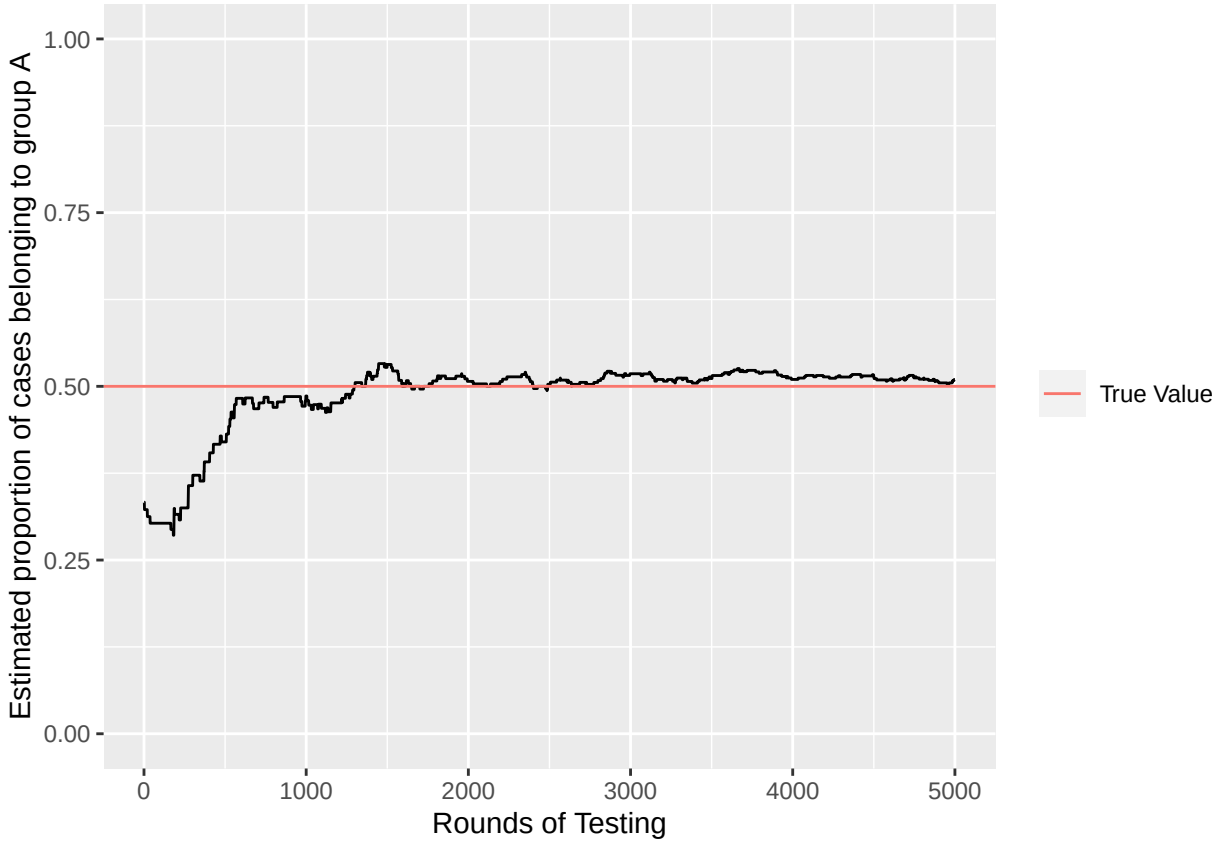
The issue with this approach is that estimating $P(Group = g|Disease = 1)$ is from observed data requires some cleverness. An analyst who forgets to account for sampling error might report estimate this quantity based on observed cases only, giving $\hat{\Lambda}_g = \frac{n_g}{n_A + n_B}$. Thus, a naive doctor will adjust their probability of followup by $\hat{Lambda_g}$ in an effort to "be Bayesian" regarding their belief about who gets the disease.

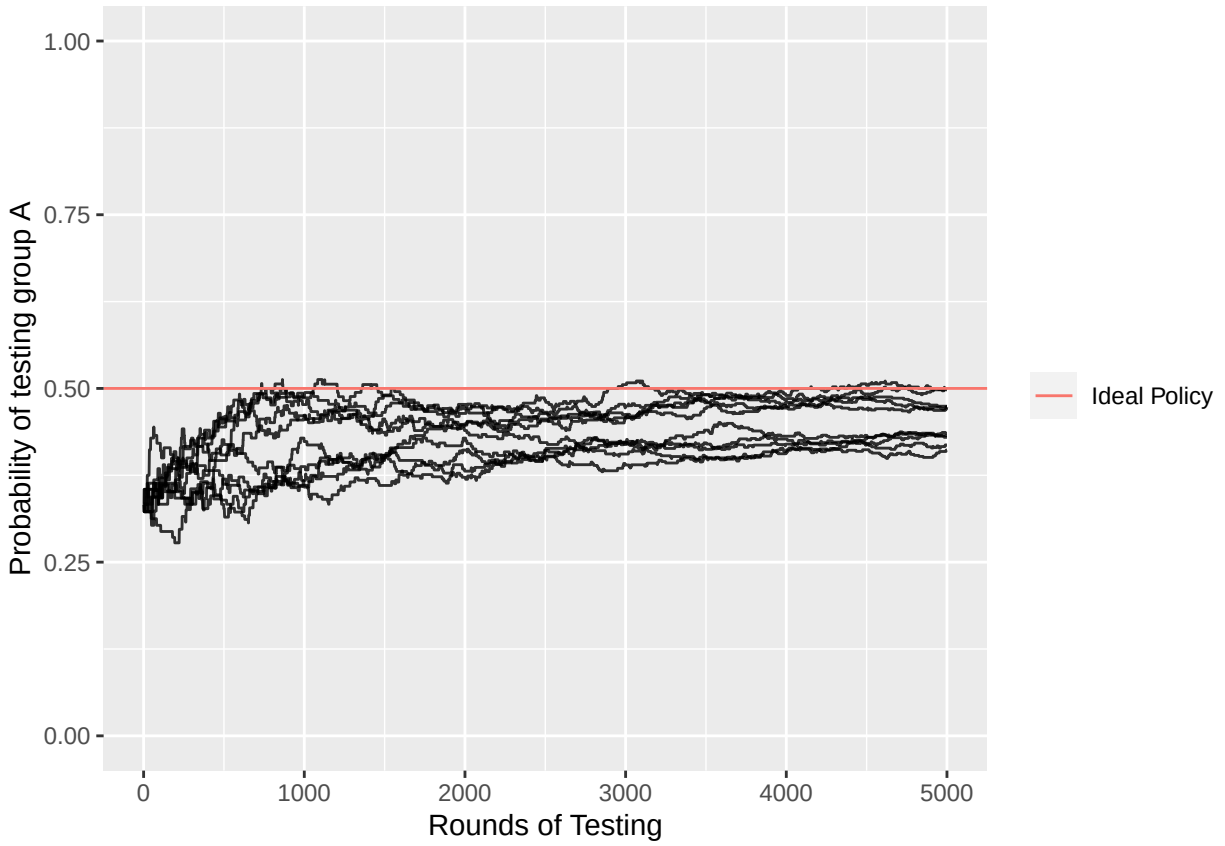
Instantaneous feedback

This first set of models will consider the behavior of a system in which diagnosed cases are immediately used to update the decision-making model for the next round of testing. Subsequently, we'll consider a system in which updates are made in batches - that is, many diagnoses occur before analysts update doctors' beliefs on how testing decisions should be made.

Case 1: Underlying rates do not differ

For now, we'll assume $\lambda_A = \lambda_B = \lambda$, so that the underlying rate of disease is the same in both groups. An ideal policy would thus test people from both groups at the same rate. Let N_A and N_B represent discovered cases when the experiment is started. We'll start with $N_A = 10$ and $N_B = 20$ with $\lambda = 0.1$. That is, the underlying rates are the same, but we have seeded the decisionmakers with some data that makes the disease appear to be more common in group B. The plot below summarizes one example simulation.

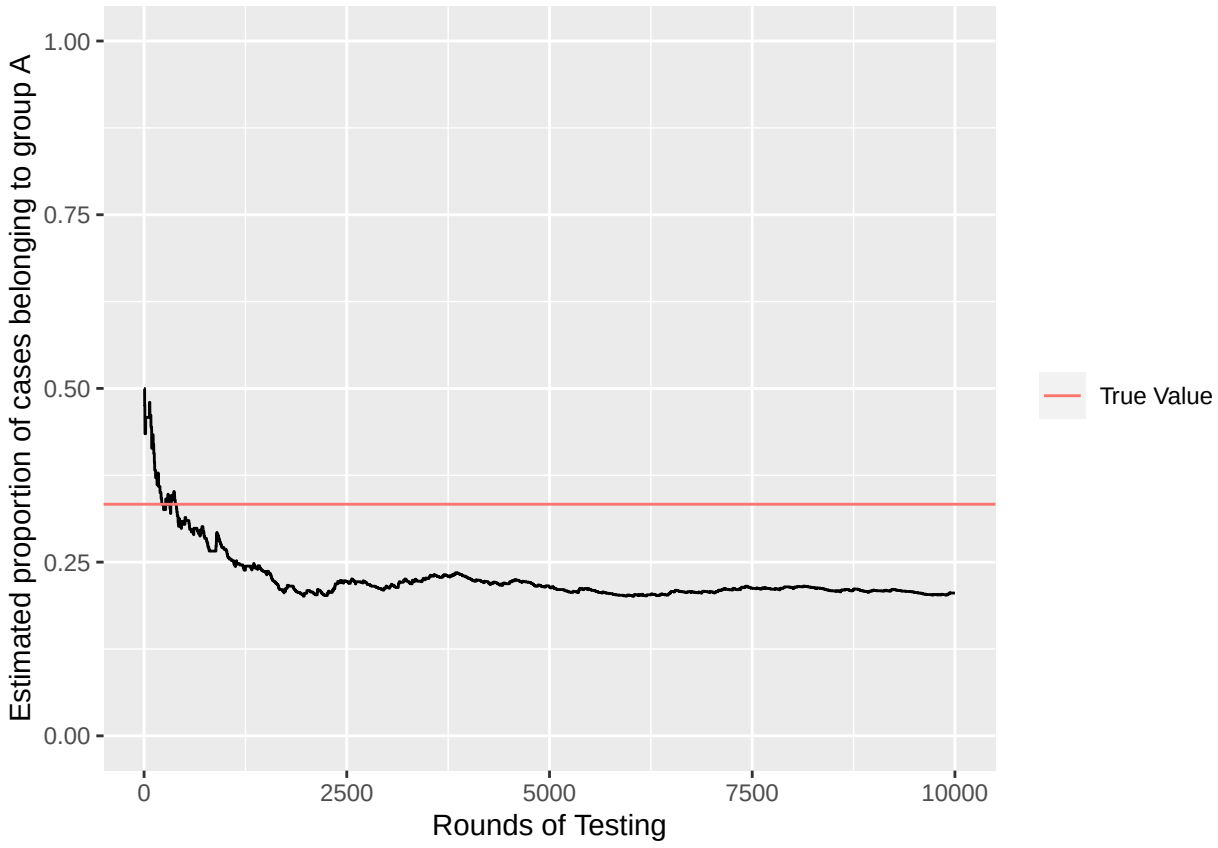


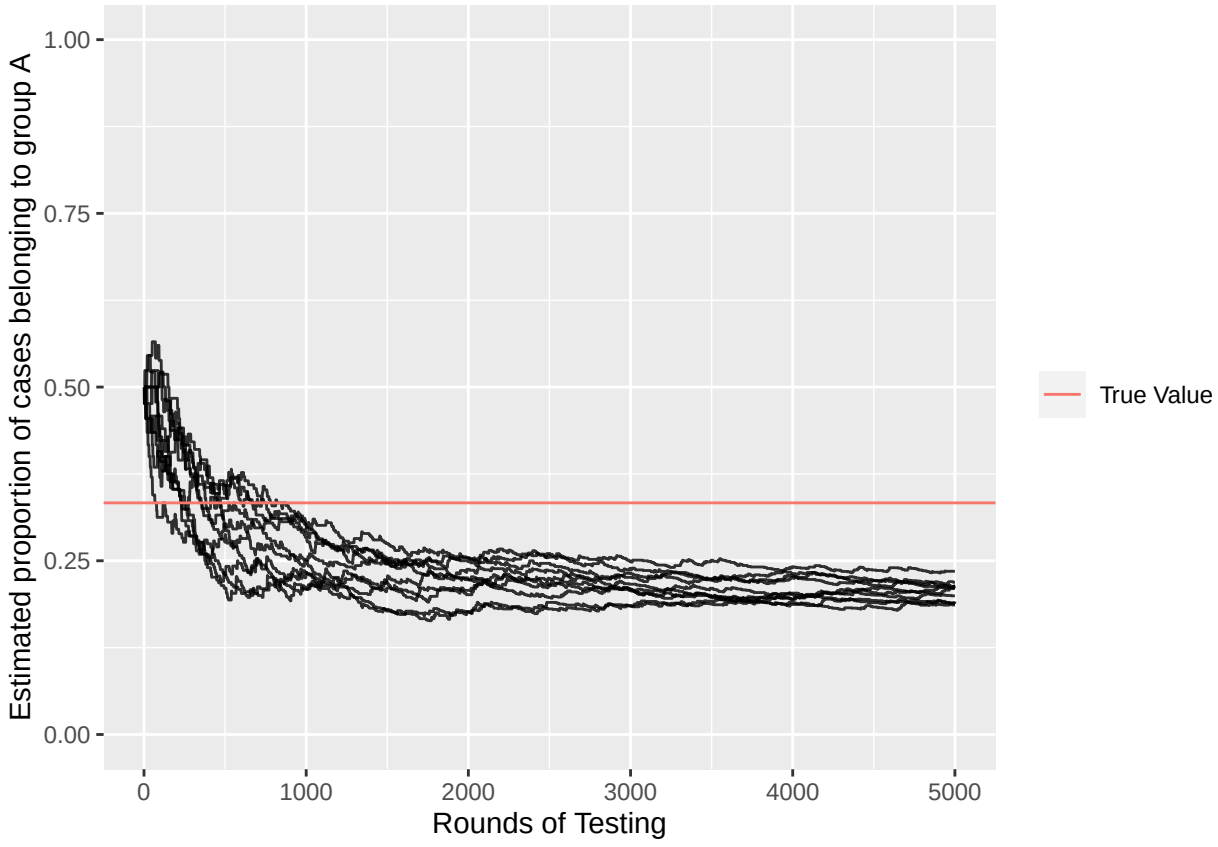


This process appears to be interestingly robust! This is a contrast to the Polya Urn decision-making set-up, which will tend to converge to an asymptotic belief about the relative rates between the two groups which is incorrect. Why the change?

Case 2: Underlying rates do differ

What happens if there is an underlying difference between the two groups - is it correctly detected? Let $\lambda_A = 0.1$ and $\lambda_B = 0.2$, so that the disease is twice as common in group B. We'll start the system off with seed data that suggests that the underlying rate of disease in both groups is the same: $N_A = N_B = 10$.


Estimated proportion of cases belonging to group A
1.00
0.75
0.50
0.25
0.00
0
2500
5000
7500
10000
Rounds of Testing
True Value


Estimated proportion of cases belonging to group A
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
5000
Rounds of Testing
True Value

Now the difference in disease rates is overestimated! By the end of many of these simulations, the doctors believe that only one in five individuals with the disease is from group A. When I run these simulation for more cycles, they seem to tend to converge around $\hat{\Lambda}_A = 0.2$. I'm not sure why the Polya Urn process tends to converge to 0 while this process tends to converge to something else...

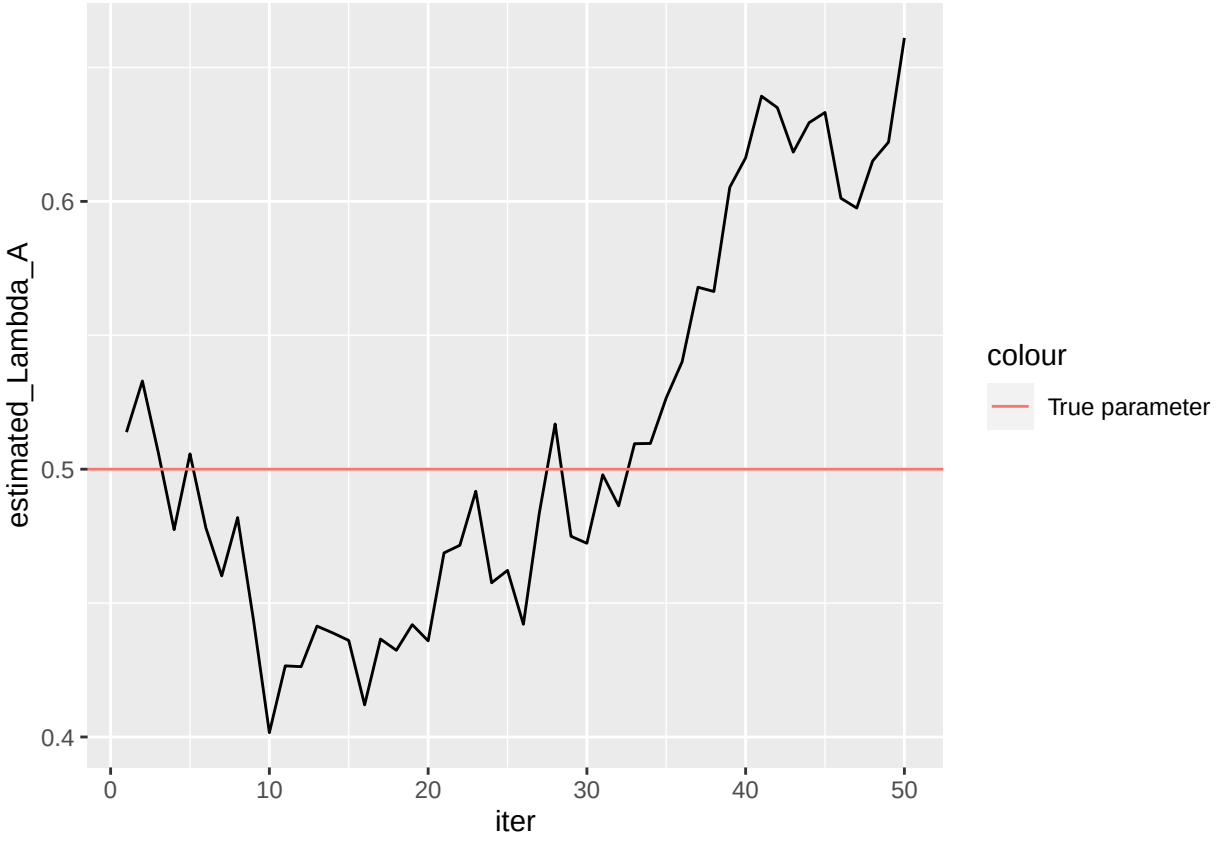
Delayed Upates

This isn't really how things tend to work in practice. Medical decision-making isn't updated instantaneously every time a test result is seen. In addition, historical data from years prior tends to be discarded in favor of more recent data, so there is a certain drop-off of old observations in favor of new ones. The next set of simulations suppose that doctors choose whom to test based on the same criteria, but they only update their beliefs about Λ_g for the groups after some number of rounds of diagnosis, n. After n rounds of diagnosis, $\hat{\Lambda}_A$ and $\hat{\Lambda}_B$ are updated with the estimates $\hat{\Lambda}_g = \frac{n_g}{n_A+n_B}$, where n_A and n_B represent the number of positive cases detected in the past n rounds of diagnosis.

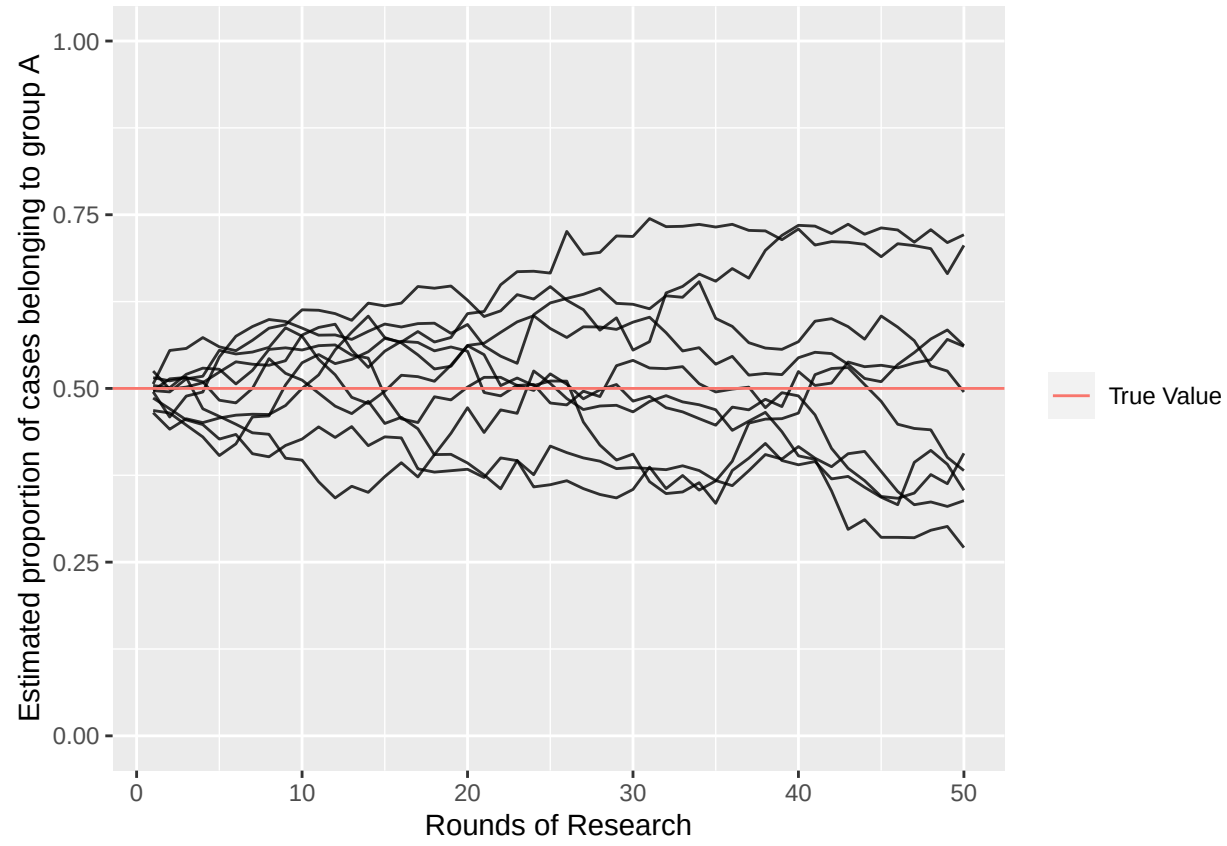
Instead of starting off with seed data, we'll now start off with a starting belief about Λ_A and Λ_B. We'll begin with agnosticism about which group bears the greater disease burden $\Lambda_A = \Lambda_B = 0.5$

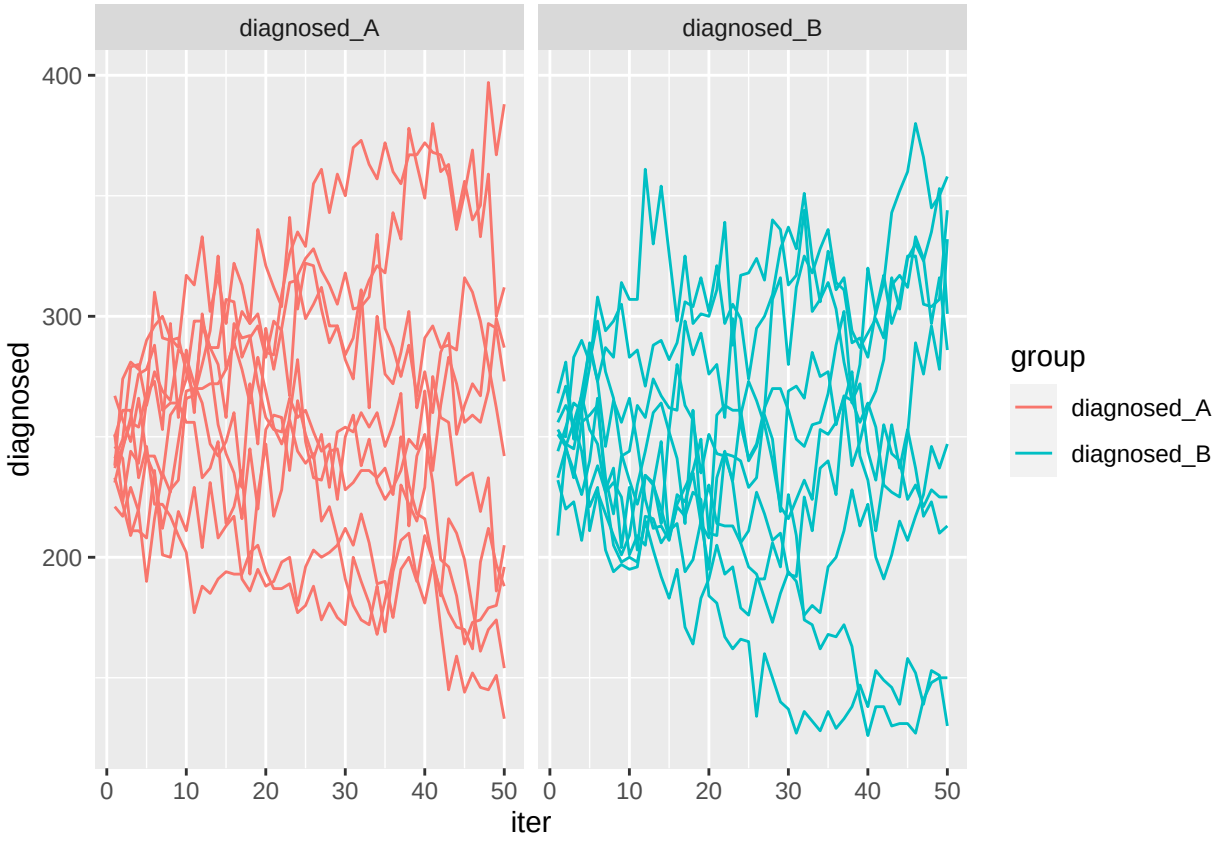
Case 1: Underlying rates do not differ.

As before, we'll assume underlying rates do not differ, and use neutral seed data. We'll suppose updates happen every 10,000 diagnoses



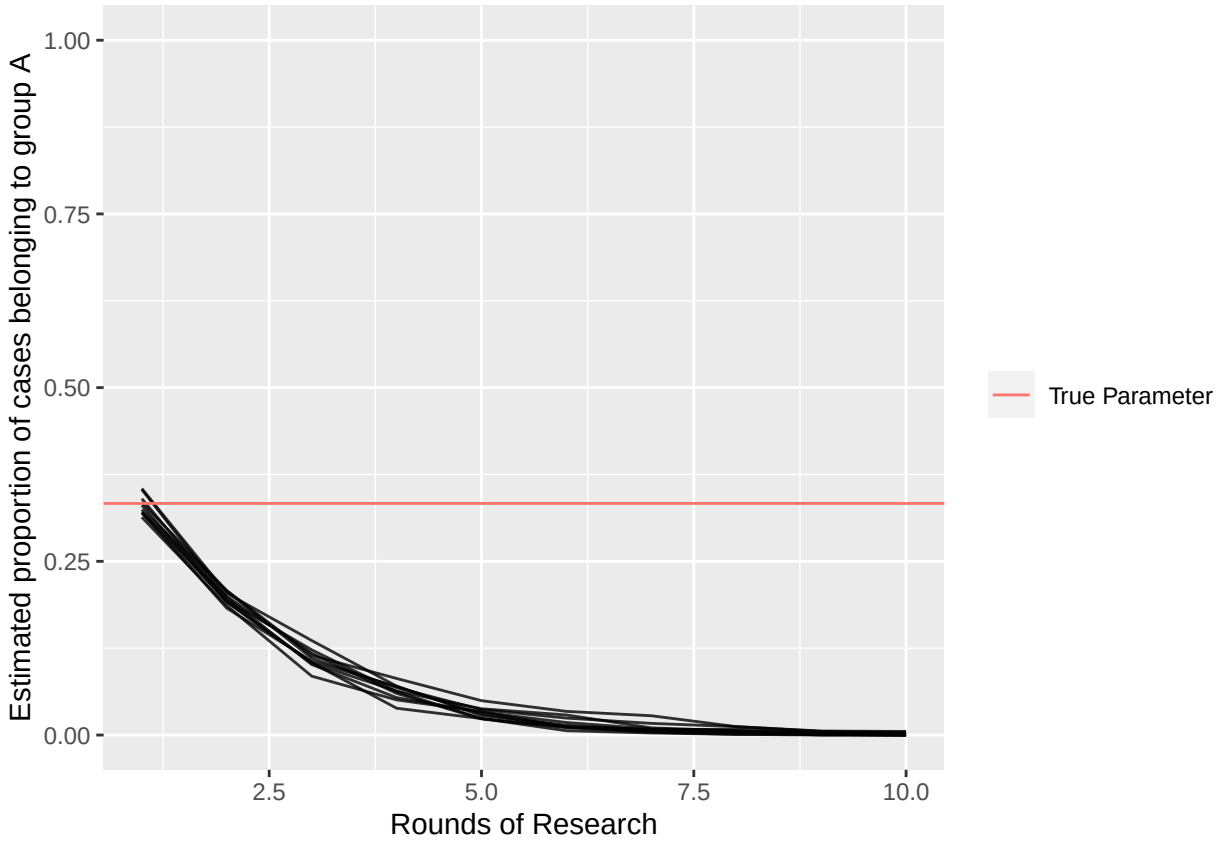
Basically it just looks like a random walk. If you run these simulations for enough iterations, they eventually get absorbed into testing one group always or the other group always. The bigger n is, the smaller the step size. Interestingly, roughly the same number of tests is run year-to-year and roughly the same number of people are diagnosed year-to-year, but as the belief randomly shifts from believing that one group has a higher prevalence than the other, one group will be tested more and diagnosed more.

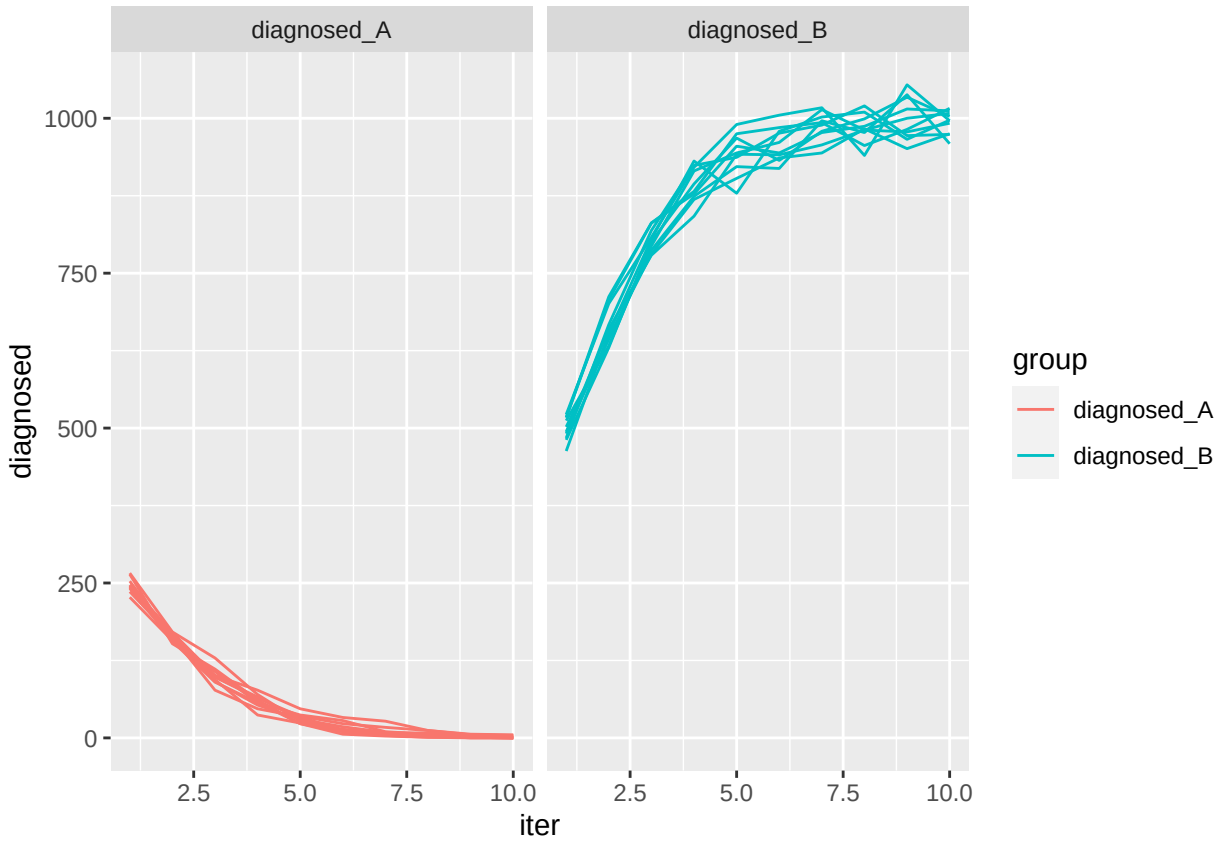

Estimated proportion of cases belonging to group A
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
50
Rounds of Research
True Value


diagnosed_A
diagnosed_B
diagnosed
400
300
200
0
10
20
30
40
50
iter
group
diagnosed_A
diagnosed_B

Case 2: Underlying rates do differ

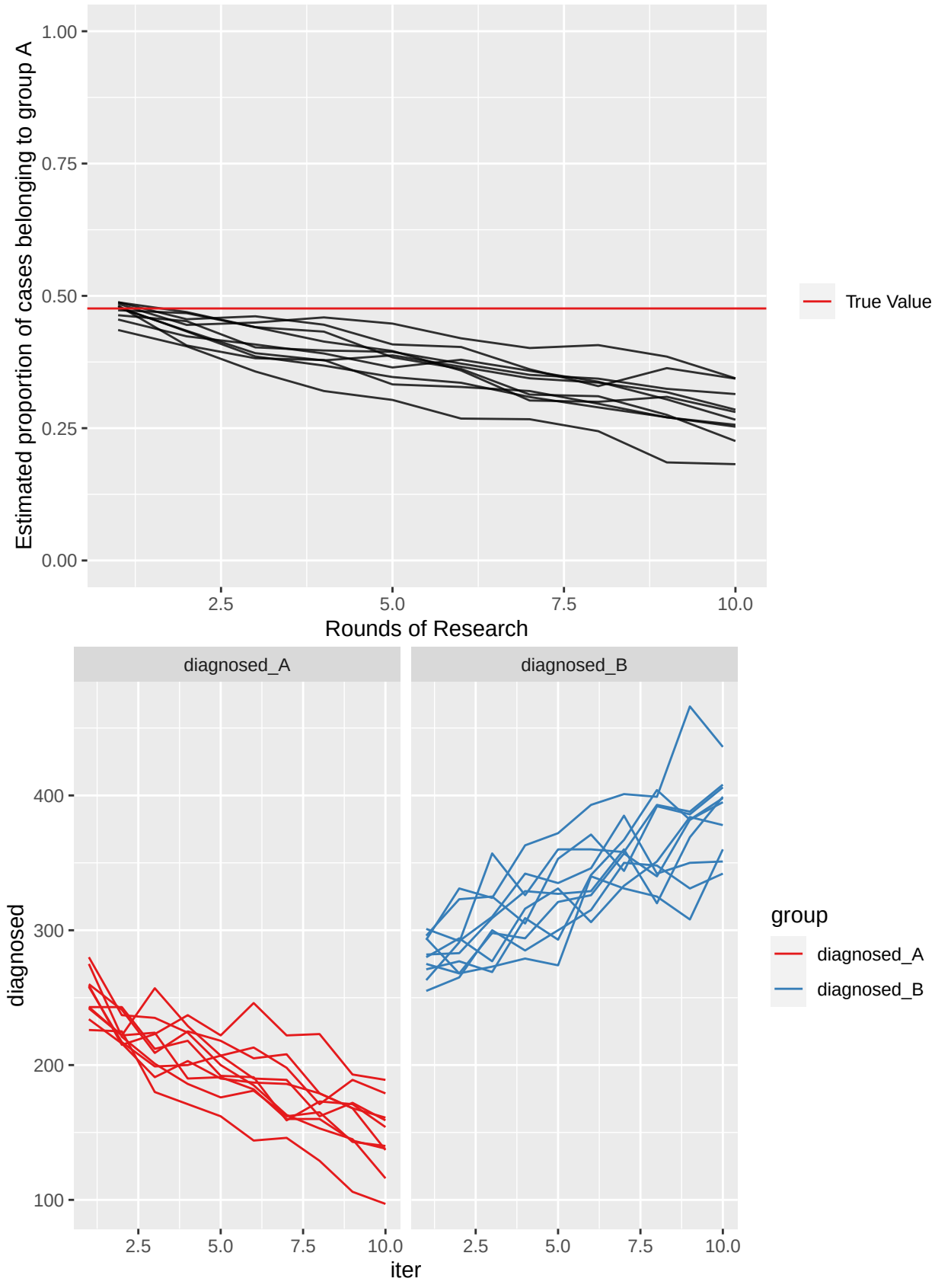
Let's take the same starting parameters, but suppose that $\lambda_A = 0.1$ and $\lambda_B = 0.2$. We'll start with the belief that the underlying disease rate is equal.





Yikes! This spirals out of control pretty immediately. The doctors in the simulation end up concluding that the disease only happens in group B. Why? First of all, in this simulation historical data is completely forgotten after every iteration. This is probably what enables this fast fixation behavior, which doesn't happen in reality.

What does the system look like if the underlying difference is smaller? The plot below shows what happens when $\Lambda_A = 0.1$ and $\Lambda_B = 0.11$, so the disease is 10% more common in group B compared to group A. The feedback loop goes out of control more slowly. After a few iterations the disease burden is overestimated quite strongly.


Estimated proportion of cases belonging to group A
1.00
0.75
0.50
0.25
0.00
2.5
5.0
7.5
10.0
Rounds of Research
True Value
diagnosed_A
diagnosed_B
diagnosed
400
300
200
100
iter
group
diagnosed_A
diagnosed_B

Concluding notes and questions

- **Prevalence equal, seed data equal:** Doesn't seem to recover the behavior of converging to a random point, but does seem to take an awful long time to converge to the correct policy. Why doesn't it randomly converge to some non-ideal policy in the way the Polya Urn stuff does?
- **Prevalence equal, seed data unequal:** Starts out biased, but always seems to recover to parity in the one-step version. Confusing because in the 100-step version unequal seed data appears to self-perpetuate. Why the difference? Would the 100-step version converge to parity if I ran it for longer?
- **Prevalence unequal:** Appears to converge to a point below the ideal. One thing I don't understand is why the 1-step version appears to converge to some nonzero point and the 100-step version appears to converge to always testing one group or the other. Why the difference? Is it because the 1-step version "sees" all the historical data, while the 100-step version only sees the past iteration of 100?