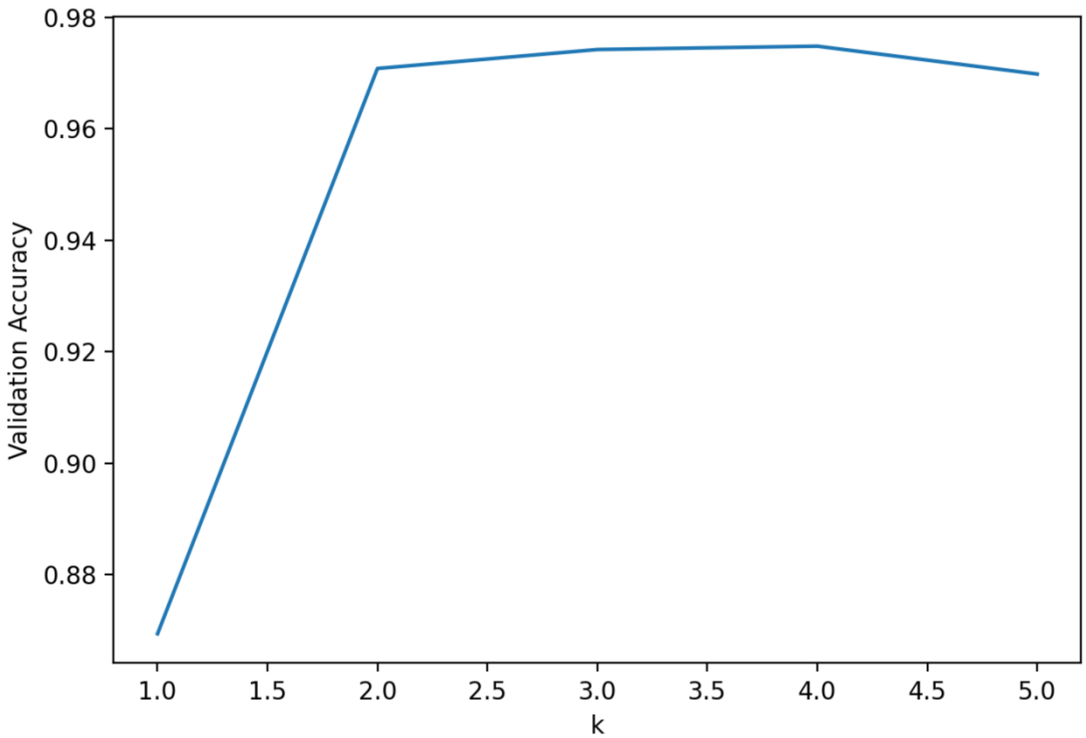
Stefan Bielmeier HW 1 Computational Genomics

2c)

Validation accuracy plot



5a) and b) Baum Welch results

This are the results for Part A and B **without** rounding the emission matrix numbers to 4 decimals.The format of the matrices is as described on the previous page.

```
Start
done, total time elapsed: 6.071180999279022
number of iterations until convergence: 67
Part A
Transition probability matrix: [[1.00000000e+00 4.06617398e-12]
 [1.00849901e-01 8.99150099e-01]]
Emission probabilities for dice 1: [0.1879  0.13825 0.16326 0.17583 0.19414 0.14062]
Emission probabilities for dice 2: [4.26946032e-18 3.53854017e-01 1.80334864e-01 5.87134385e-02
 3.30936330e-02 3.74004047e-01]

Part B
done, total time elapsed: 0.06592823266983032
number of iterations until convergence: 66
Transition probability matrix: [[9.99999999e-01 5.27480840e-10]
 [7.25176539e-02 9.27482346e-01]]
Emission probabilities for dice 1: [0.18000037 0.10900003 0.16799972 0.18899963 0.20900016 0.14500008]
Emission probabilities for dice 2: [3.47551429e-18 9.27911859e-02 3.01609274e-01 3.64723762e-01
 1.31982418e-01 1.08893360e-01]
```

These are the results for Part A and B **with** rounding the emission matrix numbers to 4 decimals (better readability)

```
Start
done, total time elapsed: 6.228426694869995
number of iterations until convergence: 67
Part A
Transition probability matrix: [[1.00000000e+00 4.06617398e-12]
 [1.00849901e-01 8.99150099e-01]]
Emission probabilities for dice 1: [0.1879 0.1382 0.1633 0.1758 0.1941 0.1406]
Emission probabilities for dice 2: [0.     0.3539 0.1803 0.0587 0.0331 0.374 ]

Part B
done, total time elapsed: 0.058191366990407306
number of iterations until convergence: 66
Transition probability matrix: [[9.99999999e-01 5.27480840e-10]
 [7.25176539e-02 9.27482346e-01]]
Emission probabilities for dice 1: [0.18  0.109 0.168 0.189 0.209 0.145]
Emission probabilities for dice 2: [0.     0.0928 0.3016 0.3647 0.132  0.1089]
```

Observations:

1. The transition probability matrix between Part A and B **does not differ significantly** for transitions **from the fair dice**. For both models, the P(fair|fair) is very close to 1, and P(loaded, fair) is very close to 0. This indicates that generally, there is a very small number of loaded dice in our model.
2. The transition probability matrix between Part A and B **does differ somewhat** (0.03) for transitions **from the loaded dice.**
    a. For Part A, approx. P(fair|loaded) = 0.1008 and P(loaded|loaded) = 0.8991
    b. For Part B, approx. P(fair|loaded) = 0.0725 and P(loaded|loaded) = 0.9275

    Generally, this indicates that once the loaded dice is rolled once, it tends to be rolled multiple times in a row. The difference hints that the first 1000 rolls contain more loaded throws than the 100,000.

3. The Emission probabilities do not differ significantly for the fair dice. This makes sense, as most rolls are fair dice rolls. Only for dice 2, we have an absolute difference of 3%.
4. The emission probabilities do **differ significantly for the loaded dice.**
   a. This is probably due to the generally low number of observations of the loaded dice being rolled in the entire sequence. In a dataset that is only 1/100th of the size of the original one, this number is even lower. Our emission probabilities heavily depend on the counts of rolls of the loaded dice. So:
      i. this may hint **an inaccurate emission probability matrix for the model trained on 1000** observations only.
      ii. The model trained on 100,000 **observations shows probably more accurate emission probabilities** for the **loaded die.**
5. Generalizing from Part A and B, it seems that the loaded dice is changed in a way that number of 1s appears minimally / not at all, while **2 and 6 are preferred over others, and 3** is somewhat preffered.