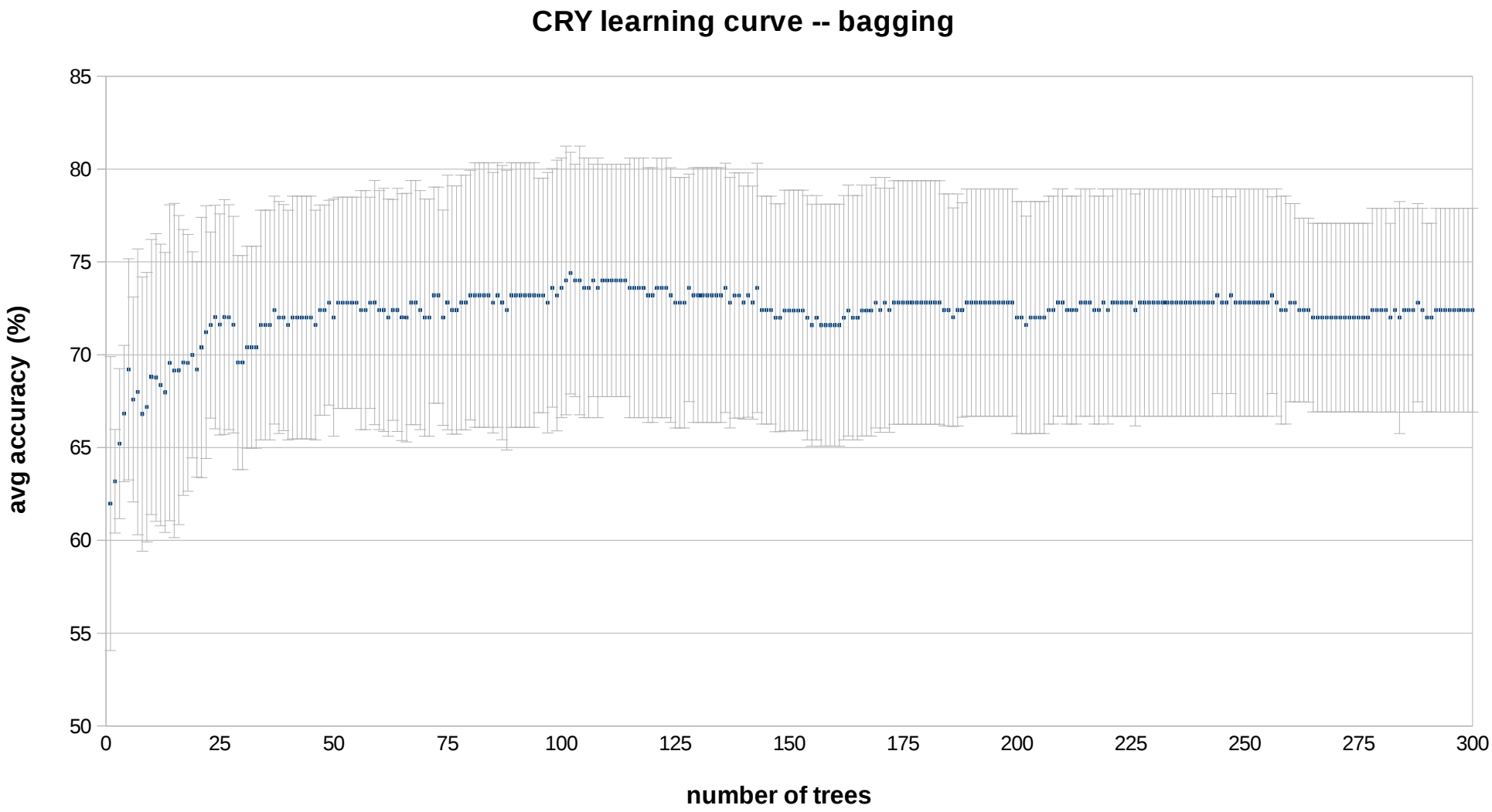
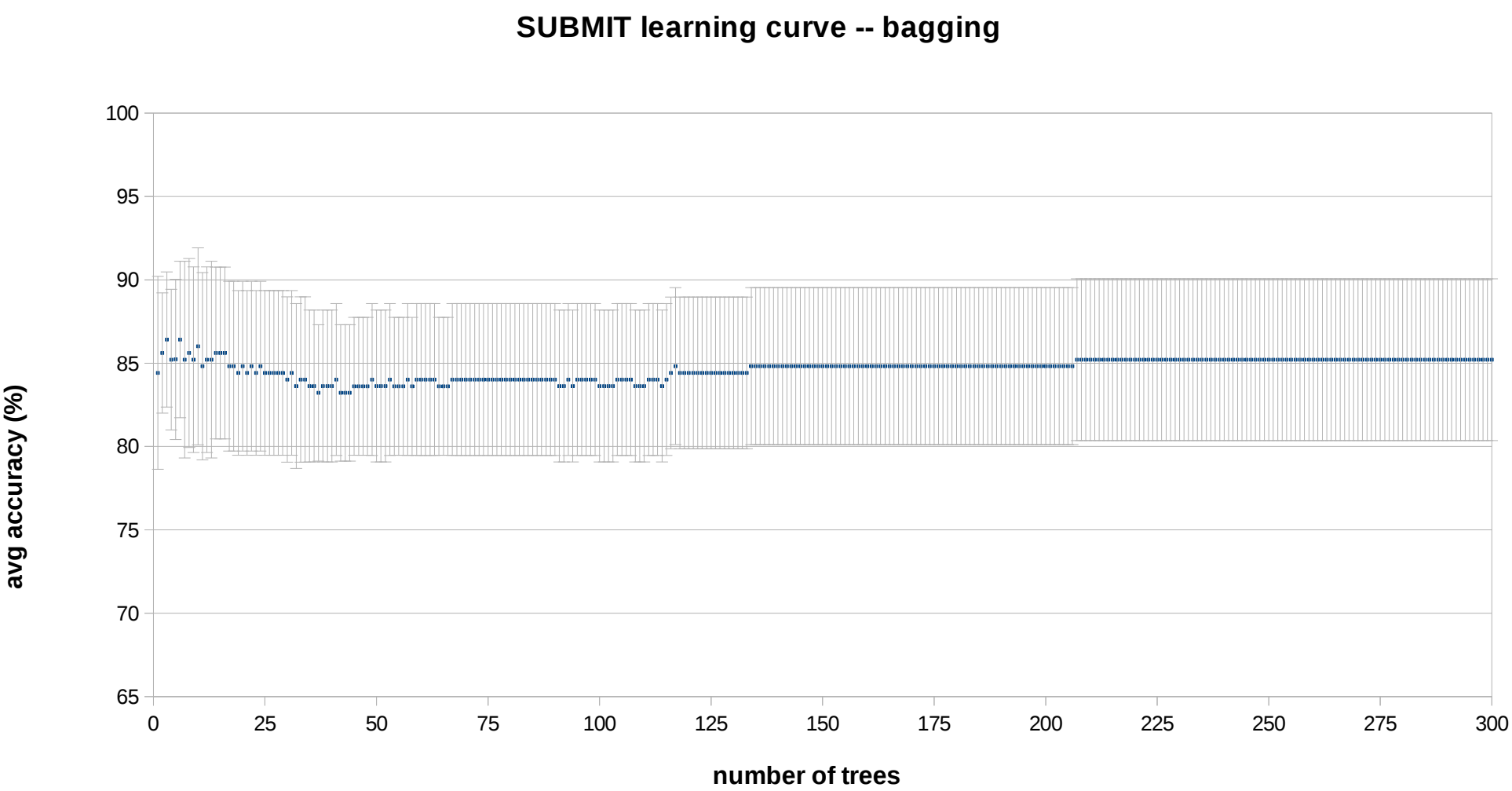
Ensembles – bagging

Comments on the illustrations below

- confidence intervals in the plots are based on 9-fold cross-validation and t-test with confidence level = 95%
- bagging is a random process, so another run would yield slightly different results
- experiment can be done using the demo code `bagging.R`

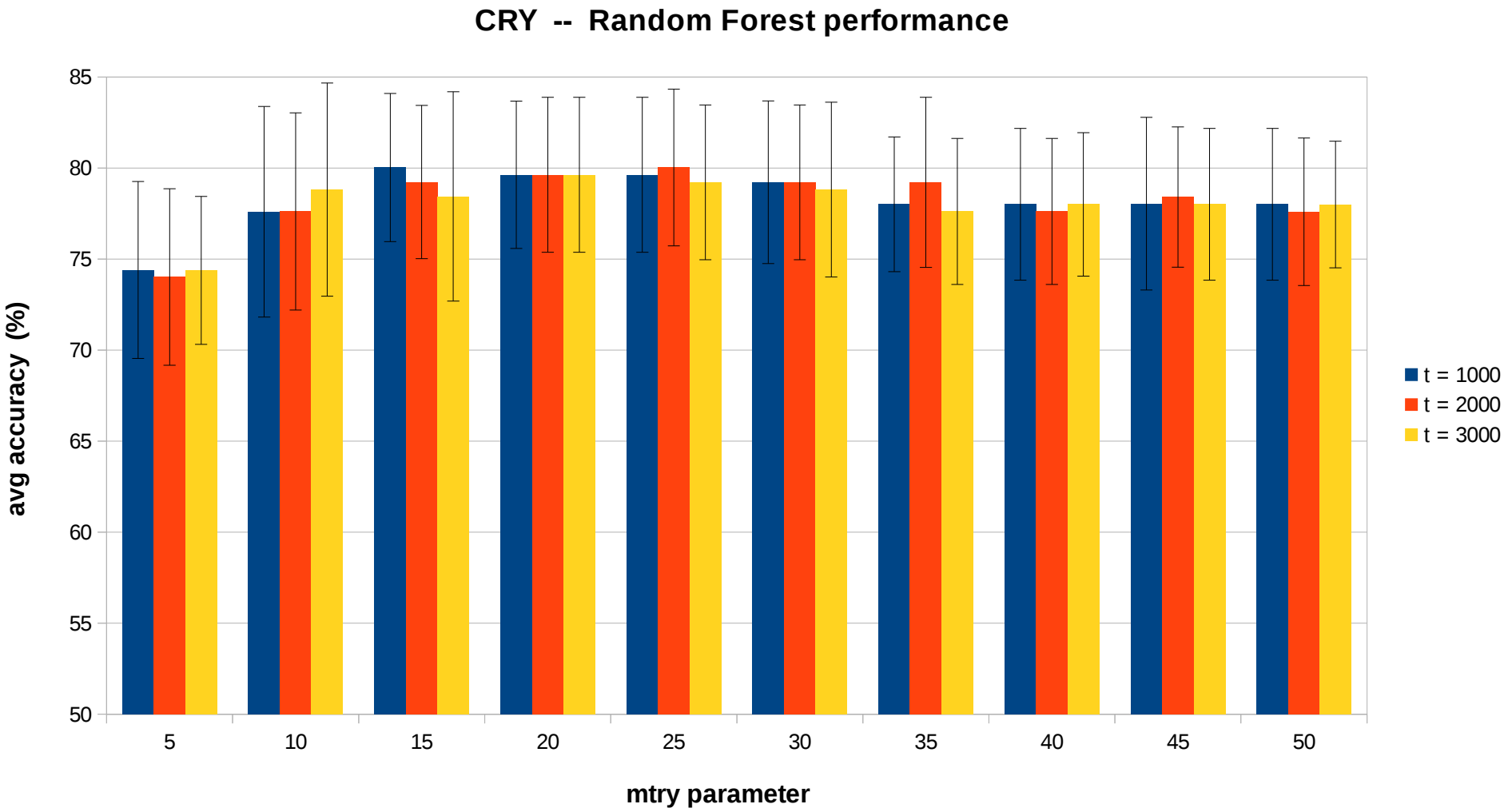

CRY learning curve -- bagging
avg accuracy (%)
85
80
75
70
65
60
55
50
0
25
50
75
100
125
150
175
200
225
250
275
300
number of trees

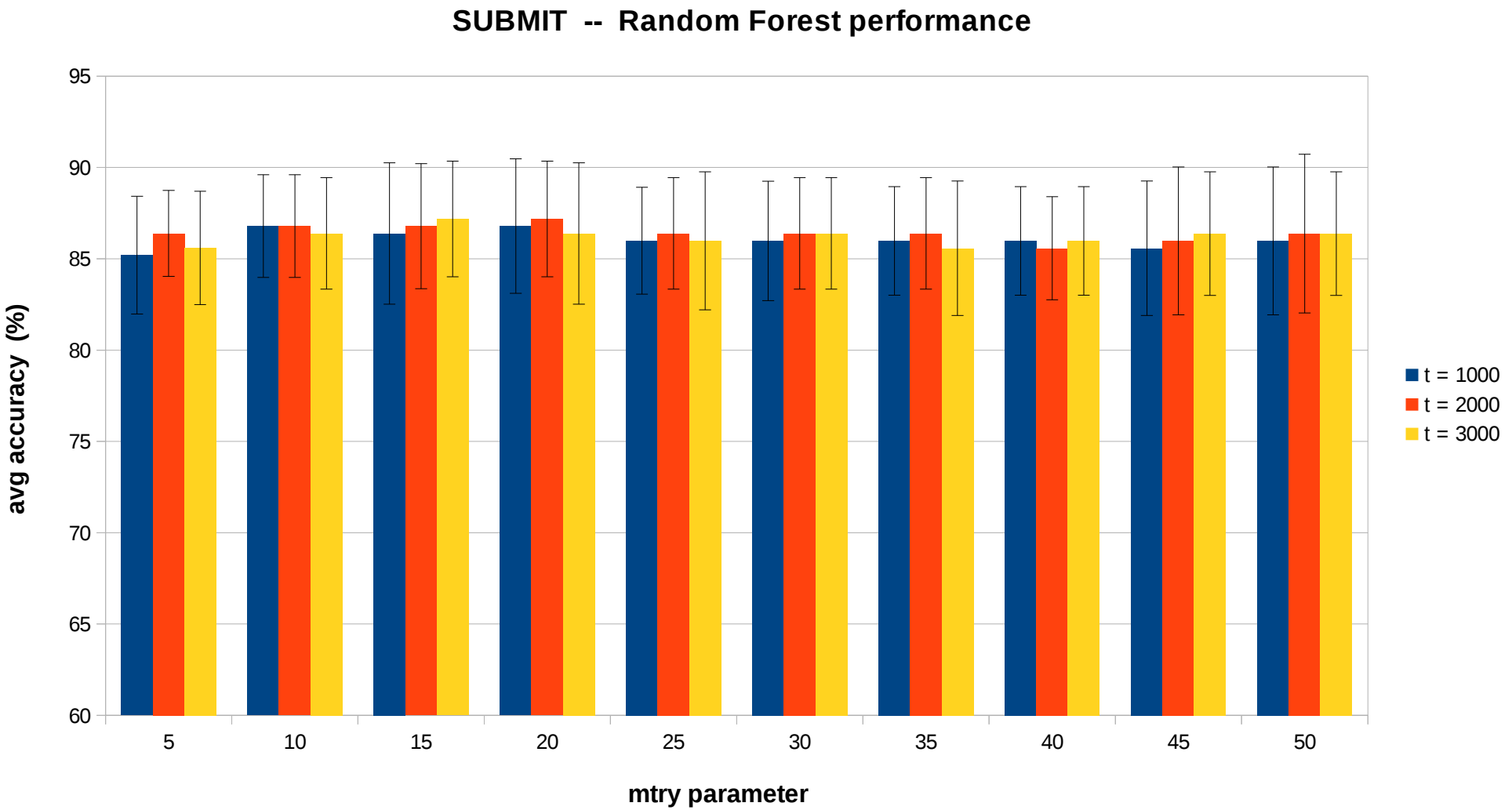

SUBMIT learning curve -- bagging
avg accuracy (%)
100
95
90
85
80
75
70
65
0
25
50
75
100
125
150
175
200
225
250
275
300
number of trees

Ensembles – Random Forests

Comments on the illustrations below

- confidence intervals in the plots are based on 9-fold cross-validation and t-test with confidence level = 95%
- mtry value means the number of the randomly sampled features when a node of a decision tree is built
- both plots show that increasing the number of trees over 1,000 does not improve the classifier performance
- experiment can be done using the demo code `rf.R`

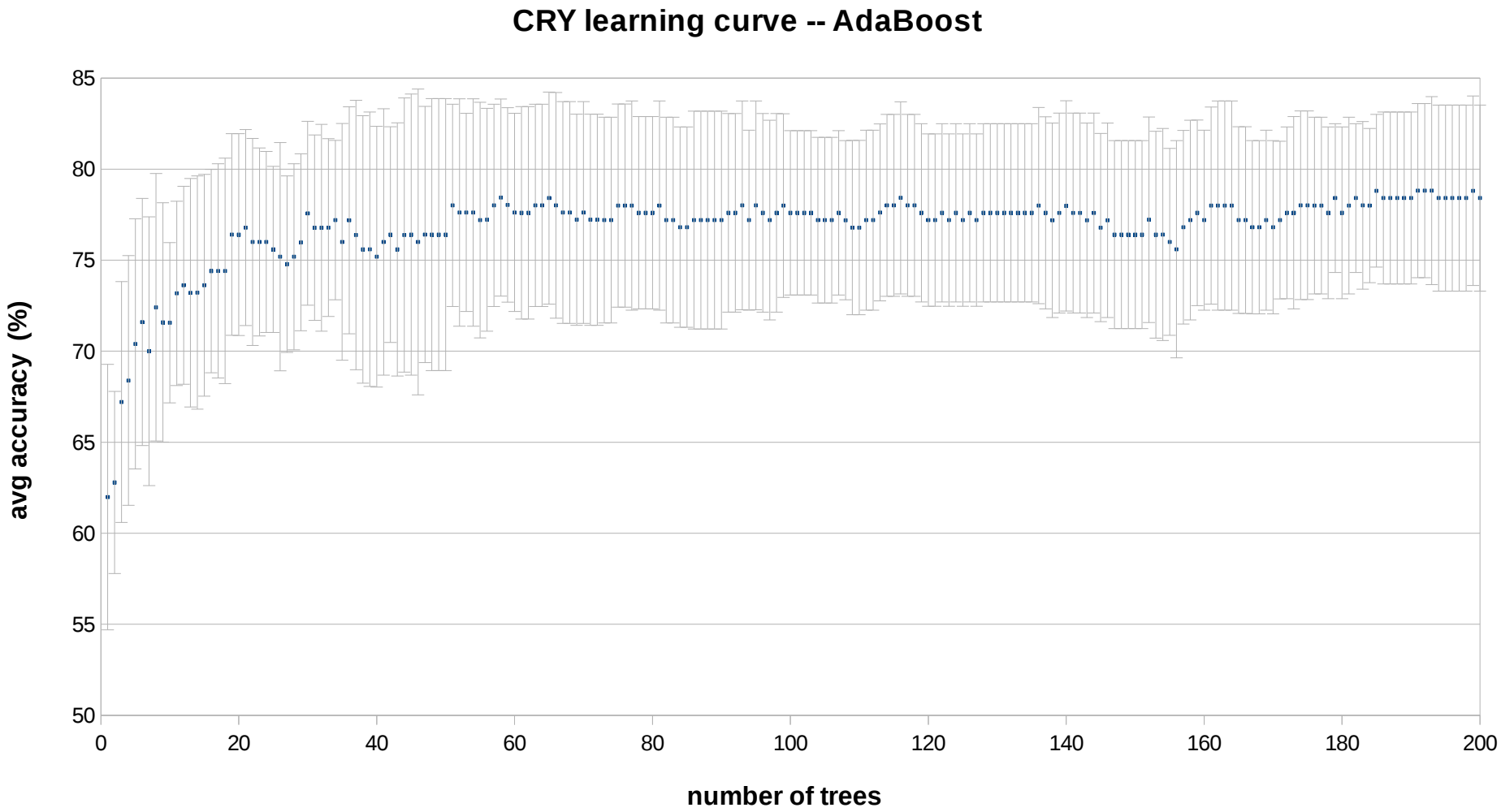

CRY -- Random Forest performance
avg accuracy (%)
85
80
75
70
65
60
55
50
5
10
15
20
25
30
35
40
45
50
mtry parameter
t = 1000
t = 2000
t = 3000

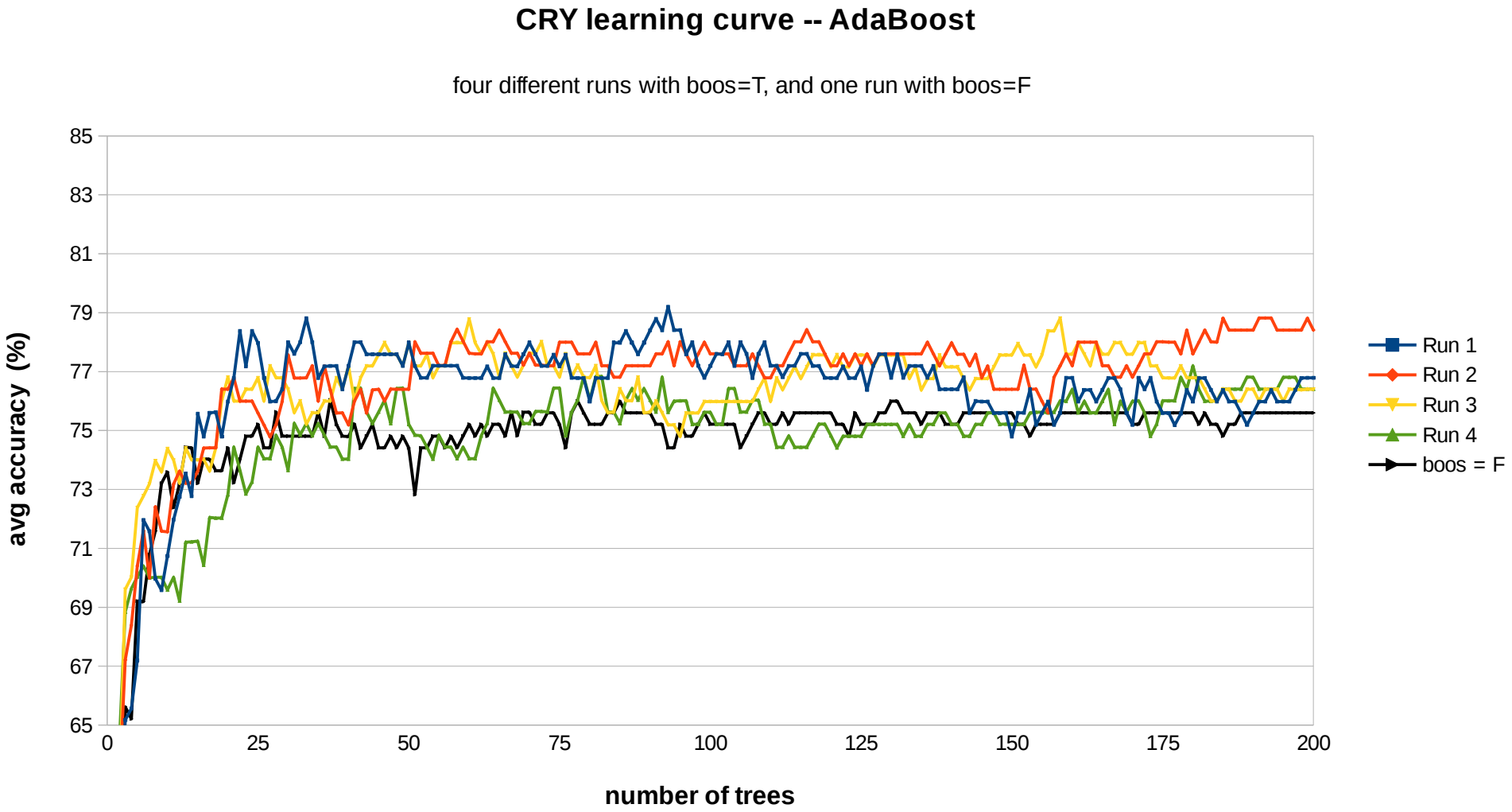

SUBMIT -- Random Forest performance
avg accuracy (%)
95
90
85
80
75
70
65
60
5
10
15
20
25
30
35
40
45
50
mtry parameter
t = 1000
t = 2000
t = 3000

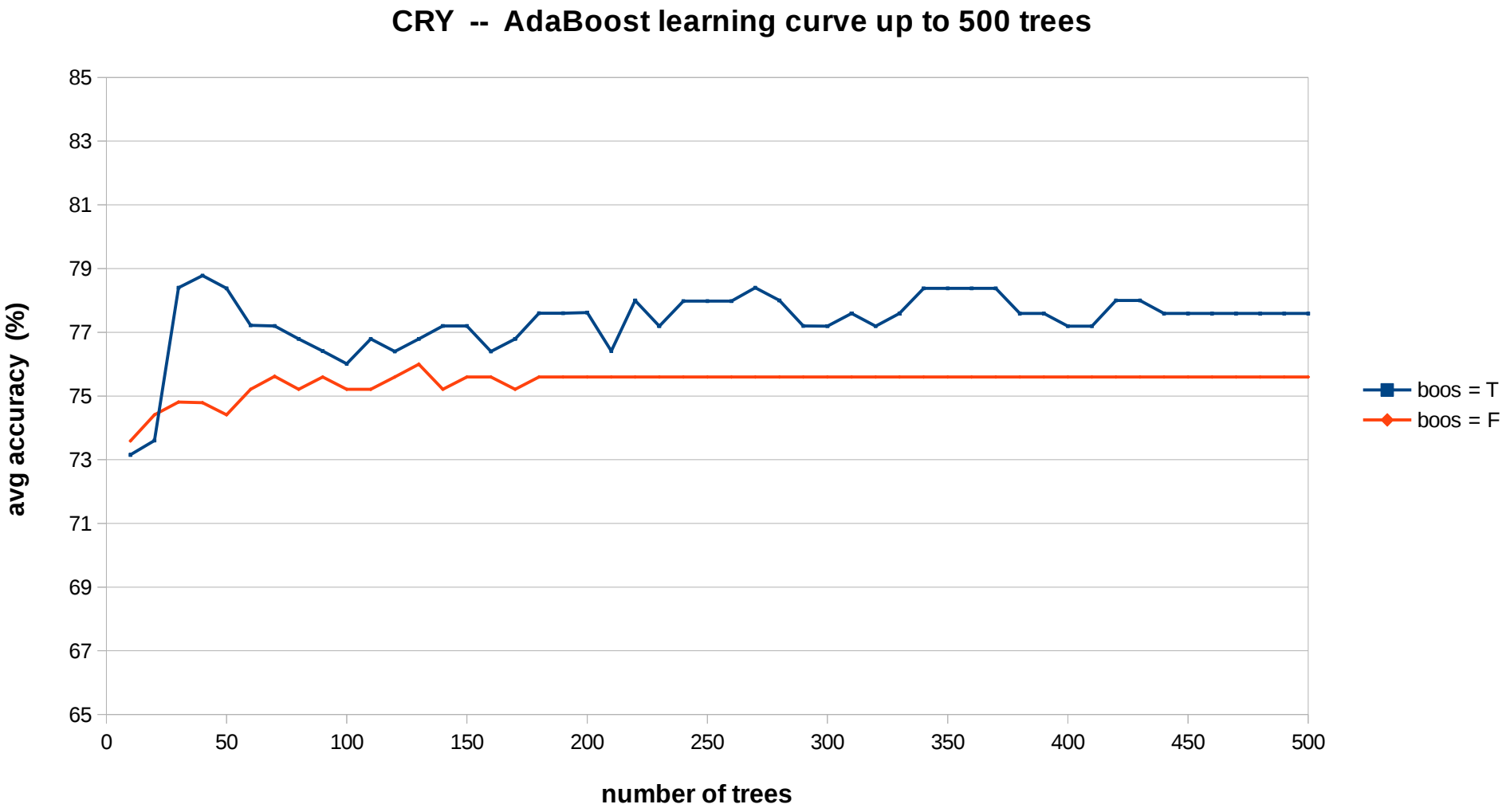
Ensembles – AdaBoost

Comments on the illustrations below

- confidence intervals in the first plot are based on 9-fold cross-validation and t-test with confidence level = 95%
- the second plot focuses on the parameter boos of the boosting procedure; it shows the comparison between random runs with boos=T and another run with boos=F
  - boos=T means that a bootstrap sample of the training set is drawn in each AdaBoost iteration; thus, the learning process is randomized and each run gives different result
  - boos=F means that no bootstrapping is done and every training instace is used with its weights
- the third plot displays the same learning procedure that goes up to 500 trees in the ensemble
- experiment can be done using the demo code `ab.R`


CRY learning curve -- AdaBoost
avg accuracy (%)
85
80
75
70
65
60
55
50
0
20
40
60
80
100
120
140
160
180
200
number of trees


CRY learning curve -- AdaBoost
four different runs with boos=T, and one run with boos=F
avg accuracy (%)
85
83
81
79
77
75
73
71
69
67
65
0
25
50
75
100
125
150
175
200
number of trees
Run 1
Run 2
Run 3
Run 4
boos = F


CRY -- AdaBoost learning curve up to 500 trees
avg accuracy (%)
85
83
81
79
77
75
73
71
69
67
65
0
50
100
150
200
250
300
350
400
450
500
number of trees
boos = T
boos = F