Making Learners (More) Monotone

Tom Viering, Alexander Mey, Marco Loog

IDA 2020

UNIVERSITY OF COPENHAGEN

TUDelft

Code available:
https://github.com/tomviering/monotone

More Data

=

Better Model

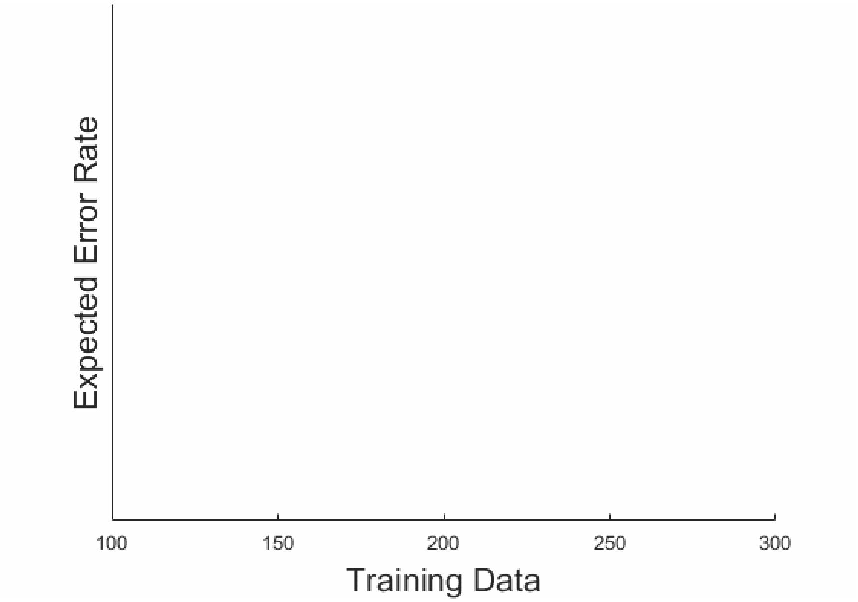
More Data

=

Better Model

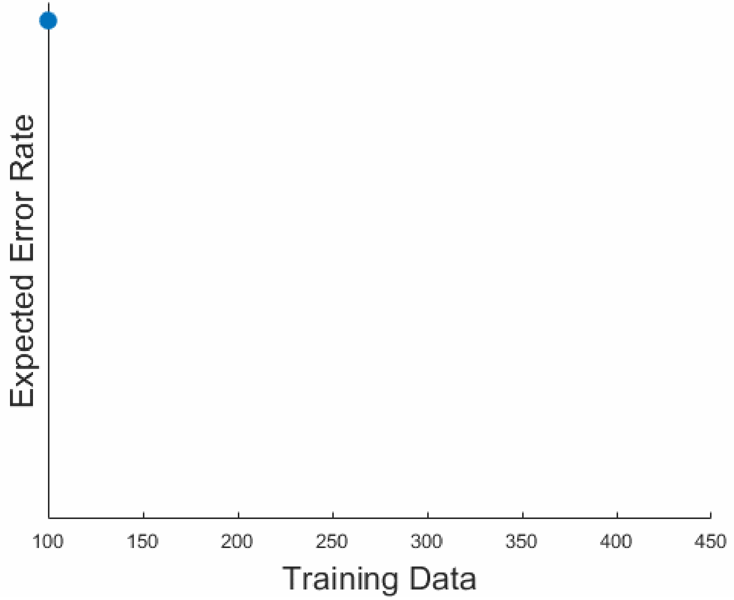
[Opper 1990], Peaking [Duin, 1995], Dipping [Loog 2012],
Double Descent [Belkin 2019], Deep Double Descent [Nakkiran 2019], Monotonicity of Learning [Viering 2019], Risk Monotonicity [Loog 2019], [Loog 2020]

Expected Learning Curve



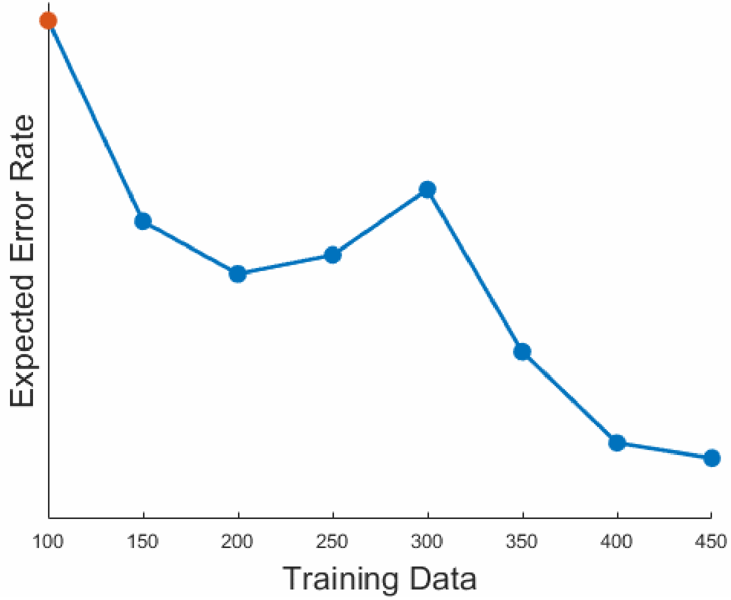
Expected = Averaged over multiple training datasets

Expected Learning Curve



Peaking Dataset [Duin, 1995]

What we want

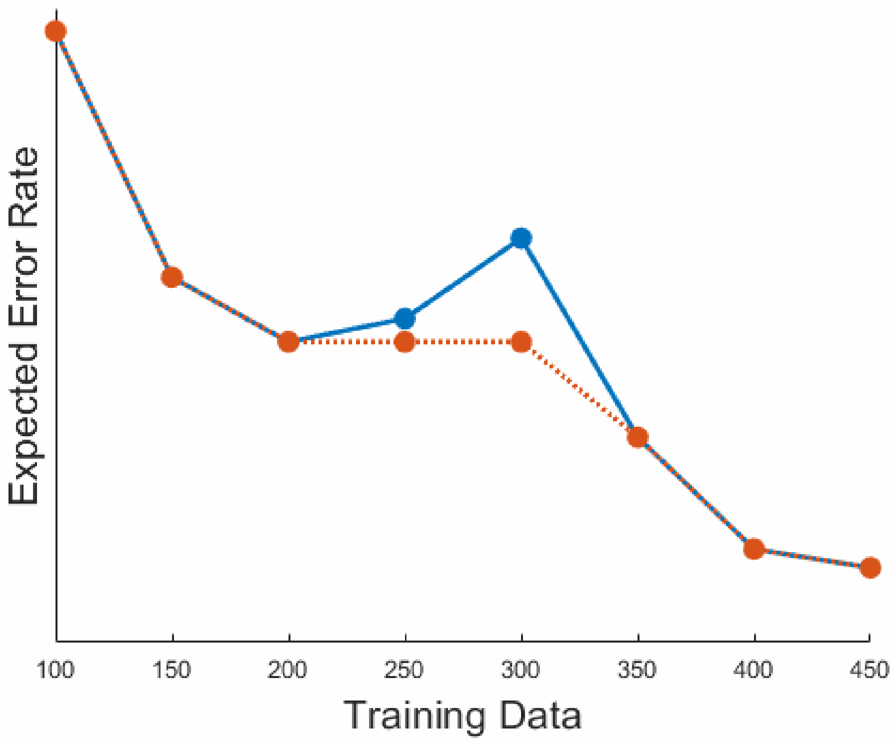


Wrapper Algorithm: makes learning curve of any classification model monotone

Wrapper Algorithm

- Two ingredients
  - Model selection
  - Conservativeness

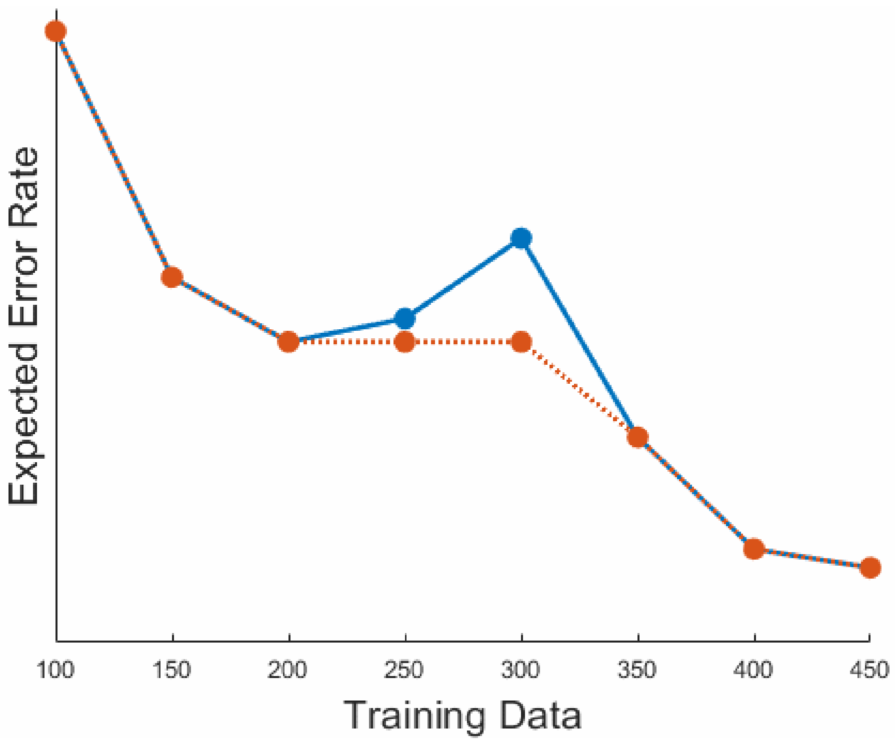
Idea 1: model selection



Pseudocode SIMPLE

For each round
  Get new data
  Train new model
  If new model better
    Use new model
  Else
    Use previous best

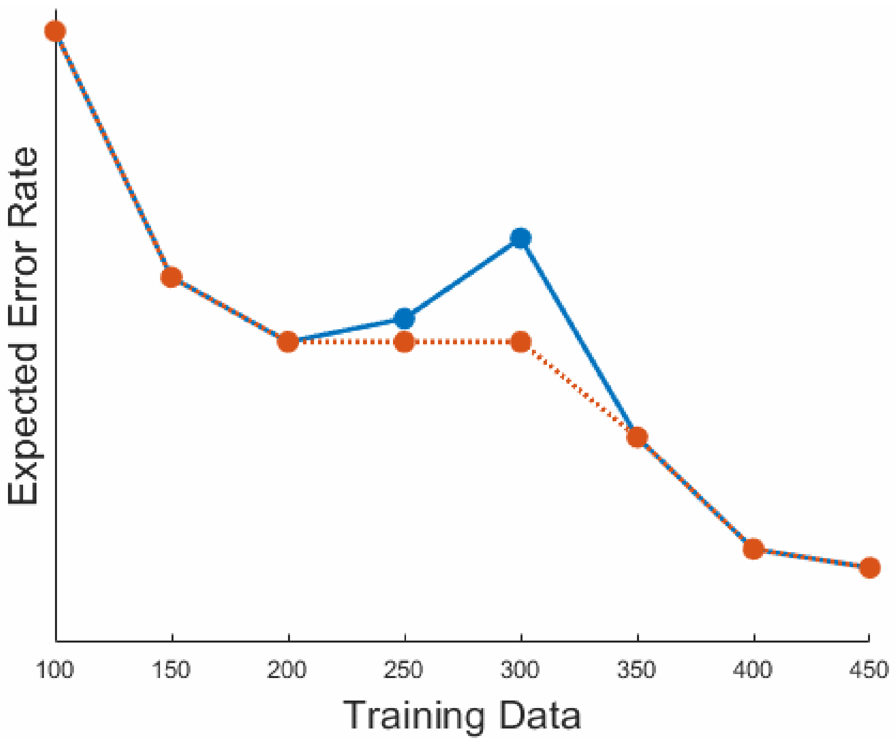
Idea 1: model selection



Pseudocode SIMPLE

For each round
  Get new data
  Train new model
  If new model better
    Use new model
  Else
    Use previous best

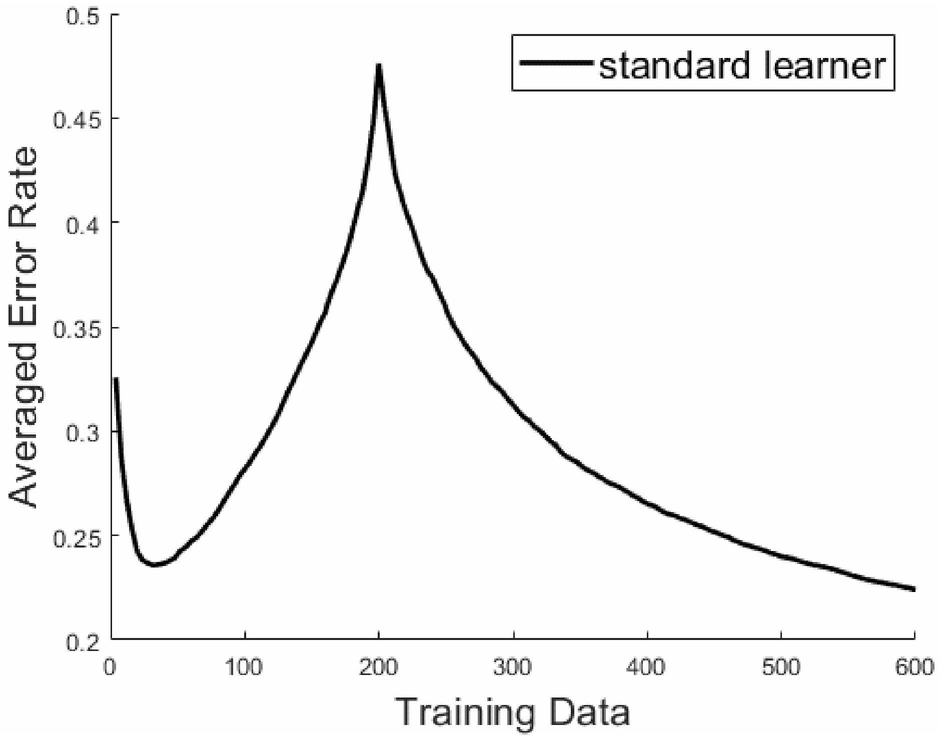
Idea 1: model selection



Pseudocode SIMPLE

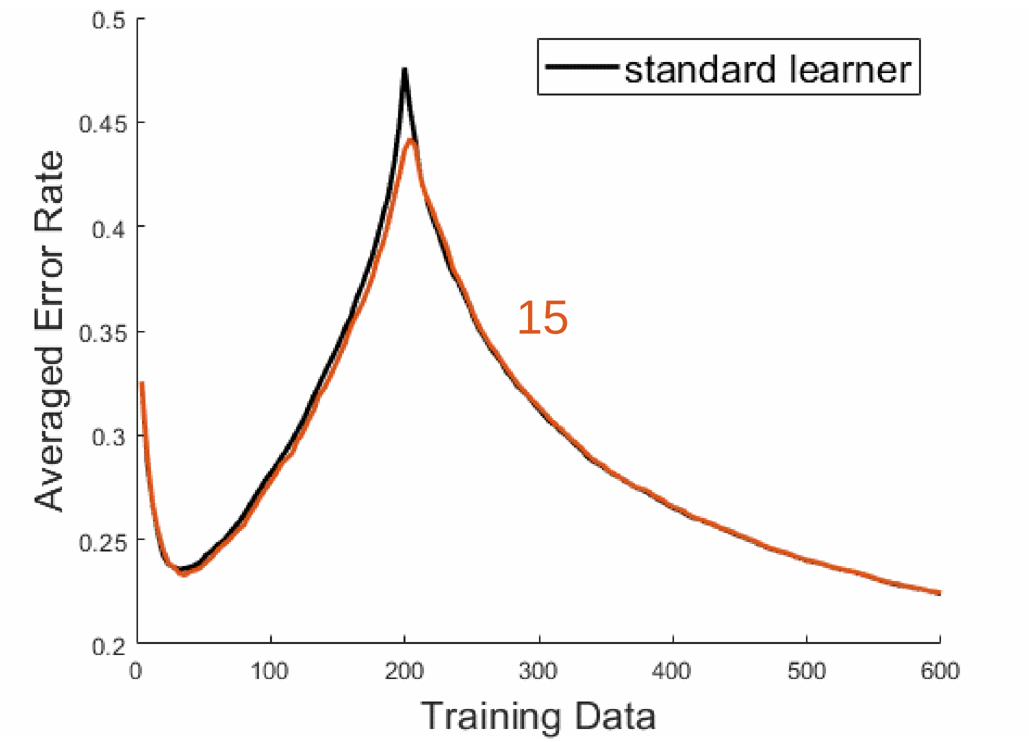
For each round
  Get new data
  Split in val, train
  Train new model
  If new better on val
    Use new model
  Else
    Use previous best

Is SIMPLE good enough?

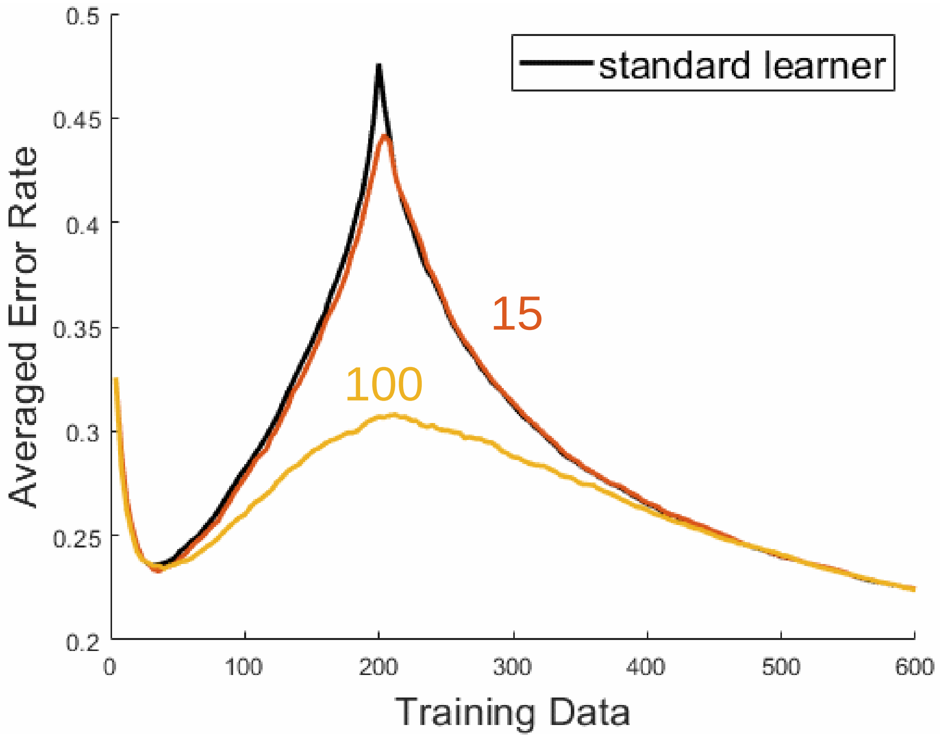


TUDelft

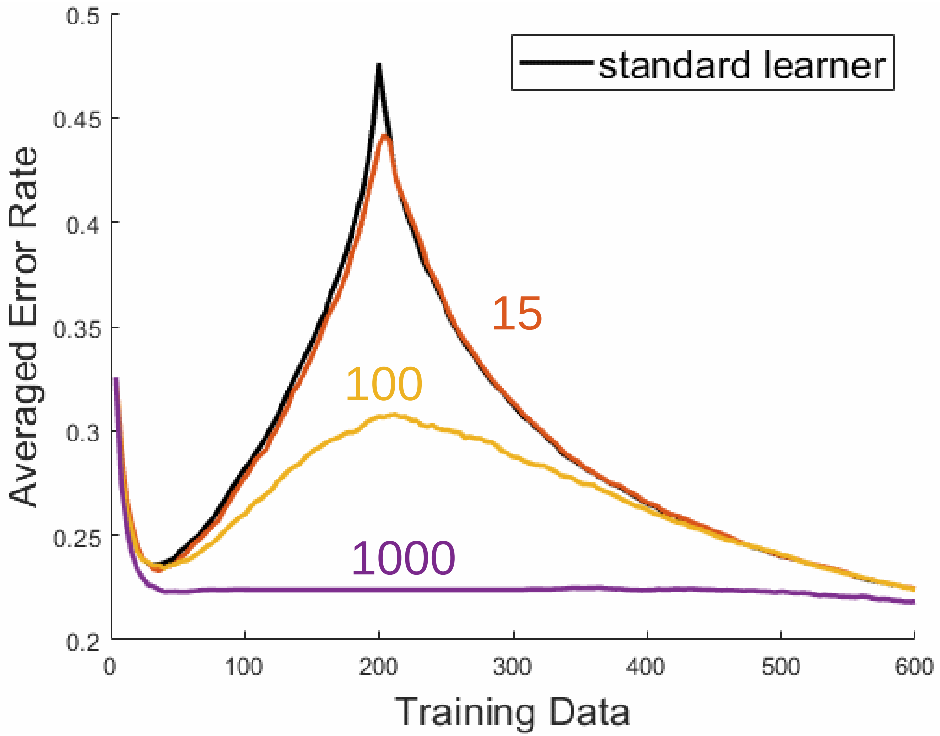
Is SIMPLE good enough?



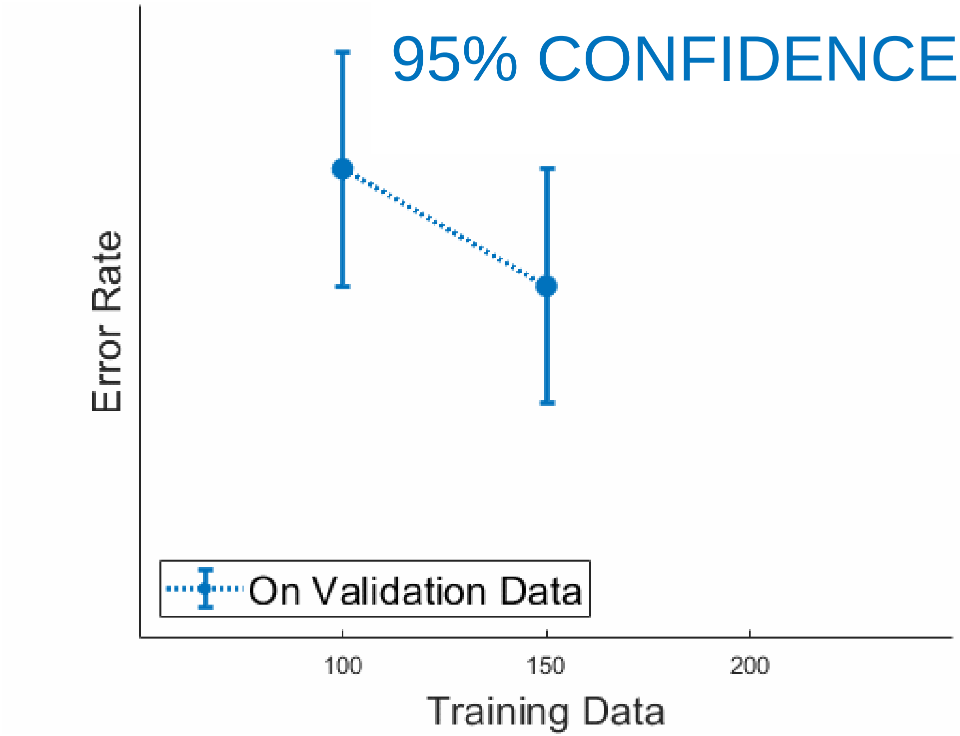
Is SIMPLE good enough?



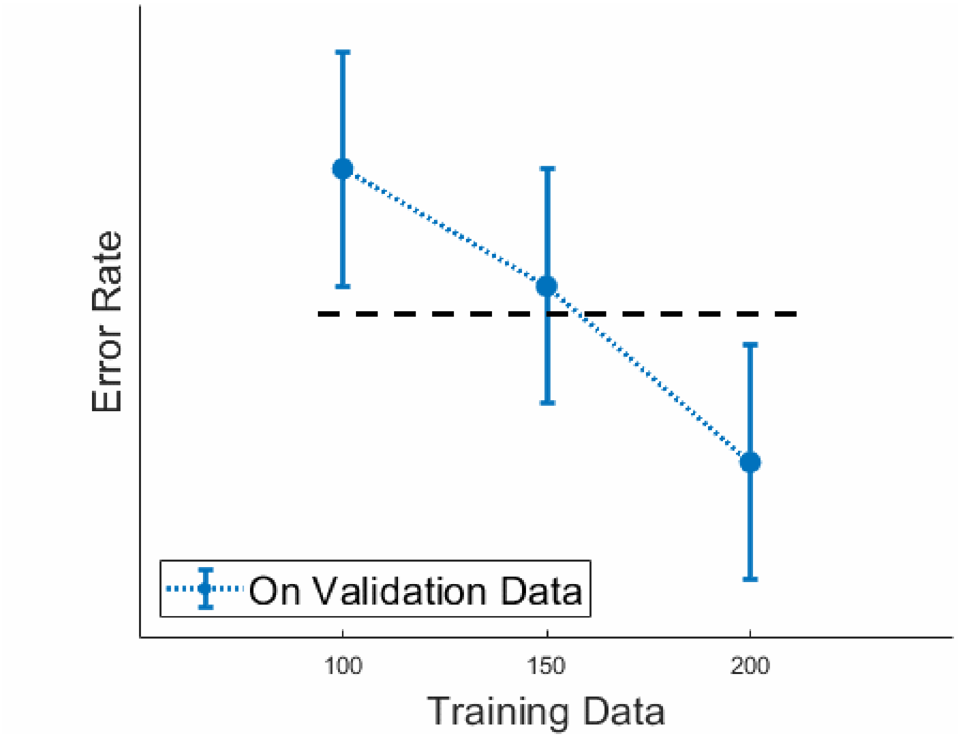
Is SIMPLE good enough?



95% CONFIDENCE INTERVAL




OOPS!!
Error Rate
True Error Rate
100
150
200
Training Data


Error Rate
On Validation Data
100
150
200
Training Data

Idea 2: Conservativeness

- Hypothesis test = conservative
  - Only switch to worse model with probability $< \alpha$

- Significance level $\alpha \in (0, \frac{1}{2}]$
  - Lower α = more conservative

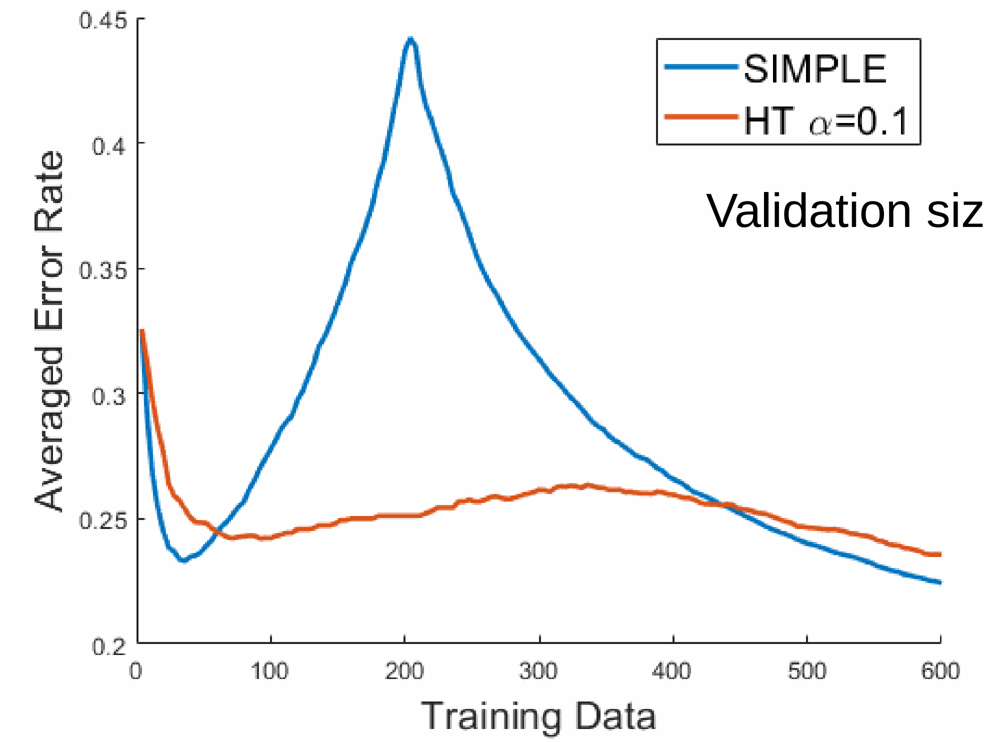
TU Delft

Theoretical Guarantees for HT

1. With probability $(1-\alpha)^n$ a single learning curve is monotone
   - Key assumption: i.i.d. data
   - Doesn't say anything about expected learning curve

2. Wrapper algorithm is consistent
   - Under some conditions…

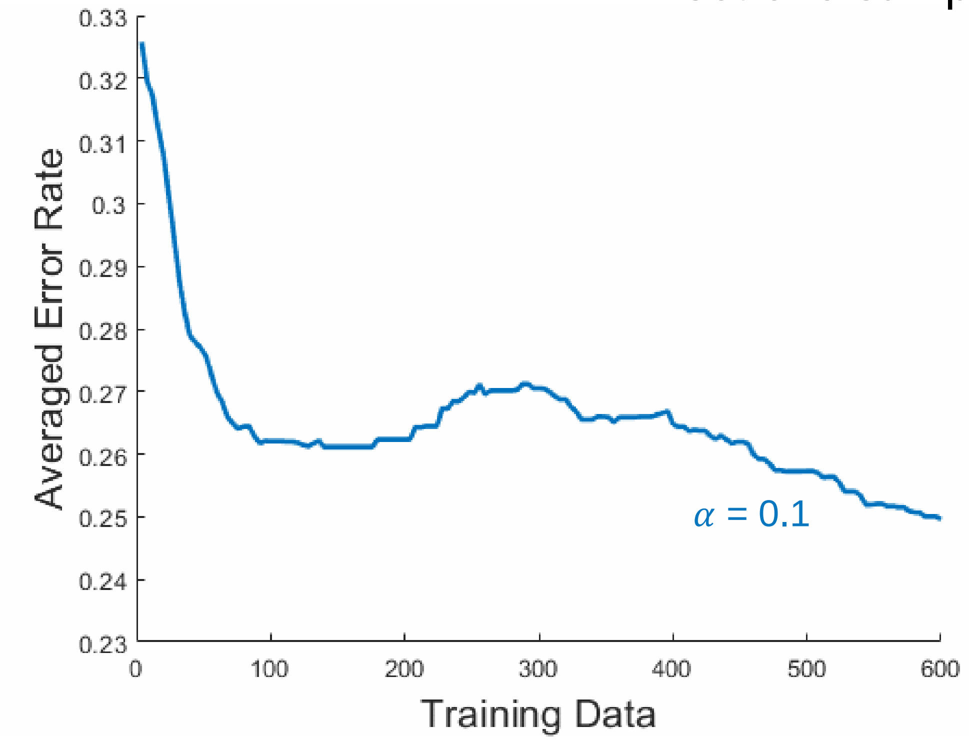
Empirical Results



Validation size the same (15)

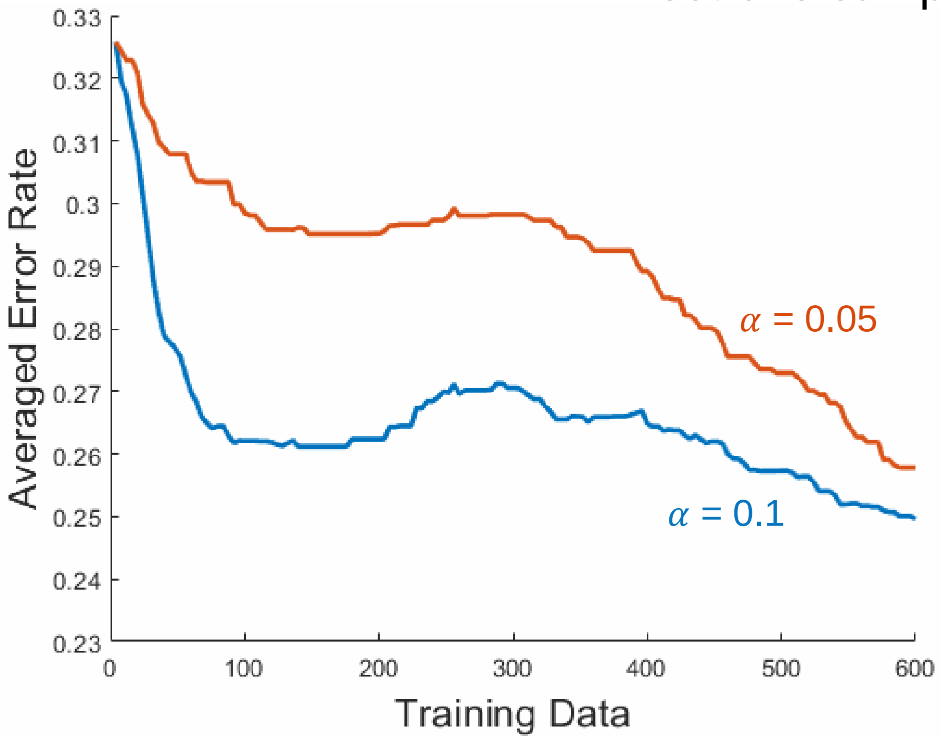
Tuning α

Very small validation set of 5 samples



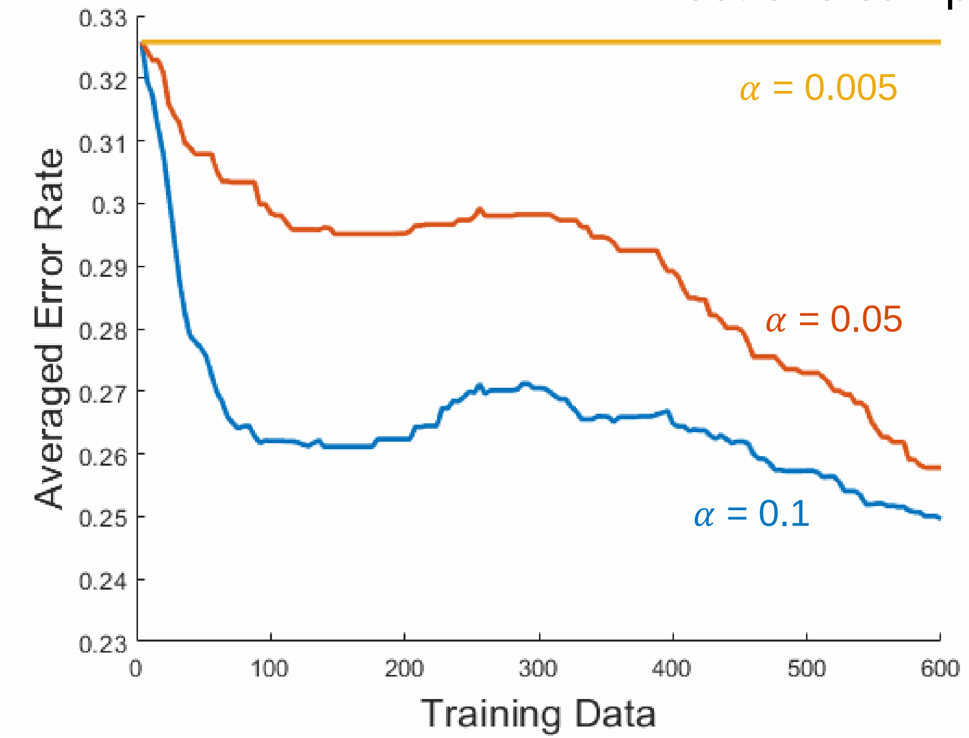
Tuning α

Very small validation set of 5 samples



Tuning α

Very small validation set of 5 samples



Benchmark

- On Peaking, Dipping, MNIST
- Several baselines

- HT is by far the most monotone
- HT is competitive in performance, but learns slightly slower
- More monotone than guaranteed

TU Delft

Discussion

- Parameter α

- Expected curve monotone?

TUDelft

Conclusion

- Make any model monotone with high probability!

- Key ingredients to achieve monotonicity
  - Model selection
  - Conservativeness

TUDelft

Making Learners (More) Monotone

Tom Viering, Alexander Mey, Marco Loog



References for non-monotone behavior:

[Duin, 1995] Small sample size generalization (‘peaking dataset’)

[Loog 2012] The dipping phenomenon

[Belkin 2019] Reconciling modern machine-learning practice and the classical bias variance trade-off

[Nakkiran 2019] Deep Double Descent: Where Bigger Models and More Data Hurt

[Viering 2019] Open problem: Monotonicity of learning.

[Loog 2019] Minimizers of the Empirical Risk and Risk Monotonicity

[Loog 2020] A Brief Prehistory of Double Descent

UNIVERSITY OF COPENHAGEN

TU Delft

Code available:
https://github.com/tomviering/monotone