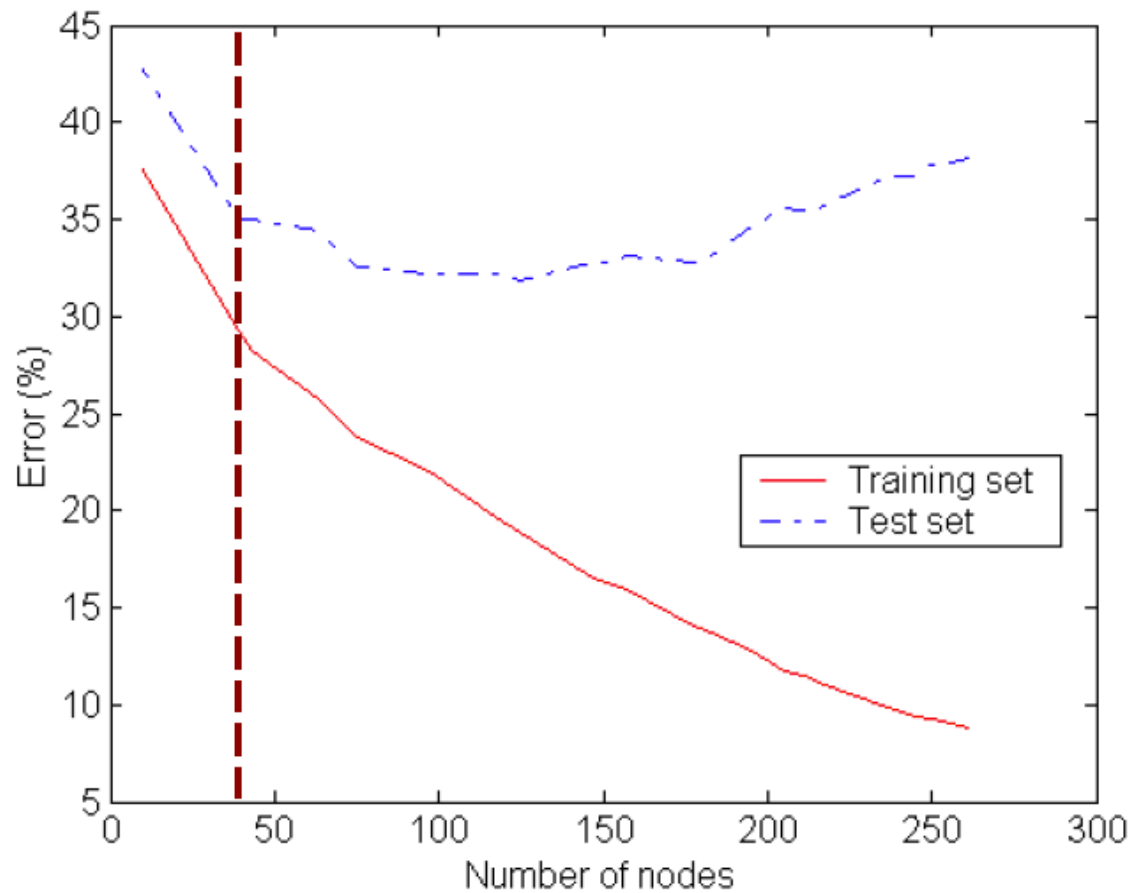


Hyperparameter Tuning

Exercise 6



Overfitting



Overfitting Task

You are optimizing a deep neural network with a huge number of hyperparameters (layers, hidden size, ..). You have a fairly large dataset, but nearly infinite computing resources. So you decide to split your dataset into a training and a test set. After an exhaustive search, you choose the combination of hyperparameters that performs best on your test set.

As soon as you apply your model in practice, you notice that the performance is considerably worse than the performance on your test set. What could be reasons for this observation?

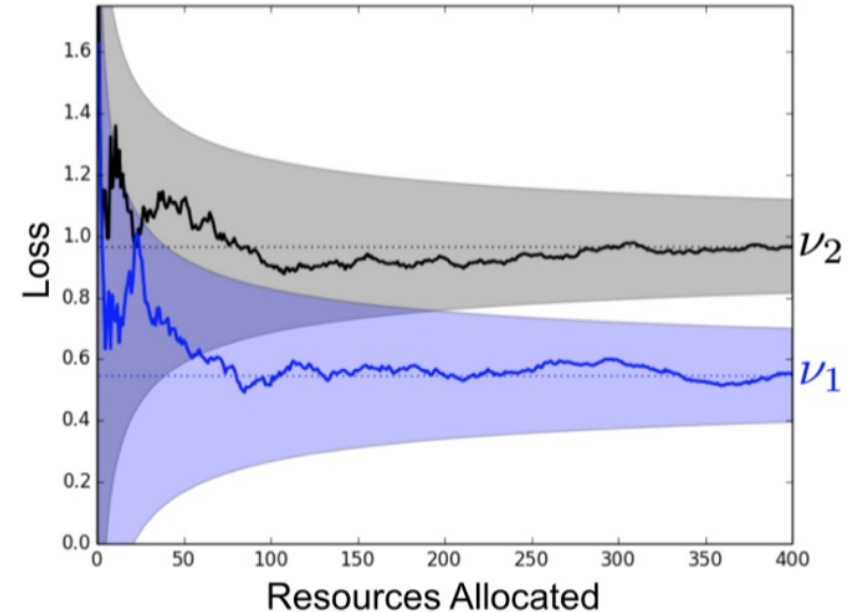
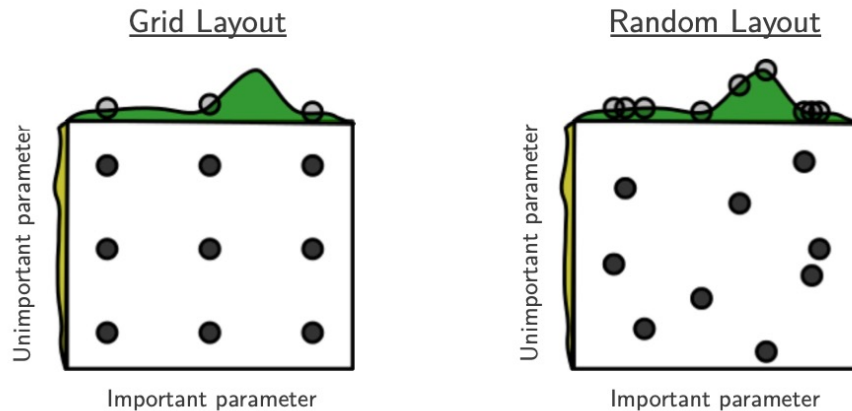
Hyperparameter Tuning

Search Strategies

- Brute Force Search
 - try out **all** hyperparameter combinations
 - computationally impossible; “blind” evaluation of parameters
- Grid Search
 - manually restrict search space to certain parameter combinations
 - quality of solution strongly dependent on grid definition
- Hill-climbing Heuristics
 - find (near-)optimal solution through evaluation of neighborhood
- Genetic Algorithms
 - find (near-)optimal solution through cross-over and mutation

Hyperparameter Tuning

Successive Halving



Successive Halving

1. Sample set of hyperparameter configurations
2. Evaluate performances of current configurations
3. Sort by performance and throw out bottom half
4. Go back to 2. and repeat until one config remains

Hyperparameter Tuning

Bayesian Optimization

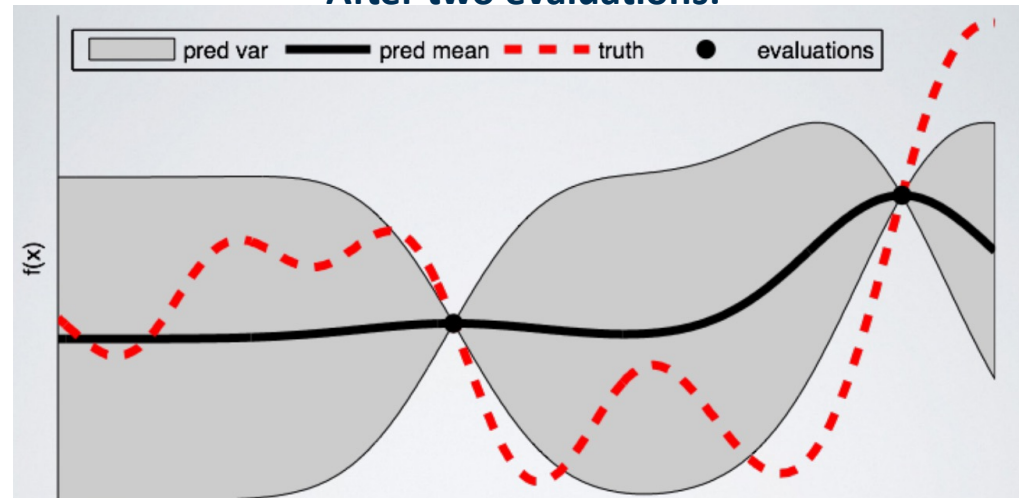
Bayesian Hyperparameter Optimization

1. Build a surrogate probability model of the objective function
2. Find the hyperparameters that perform best on the surrogate
3. Apply these hyperparameters to the true objective function
4. Update the surrogate model incorporating the new results
5. Repeat steps 2-4 until max. iterations or time is reached

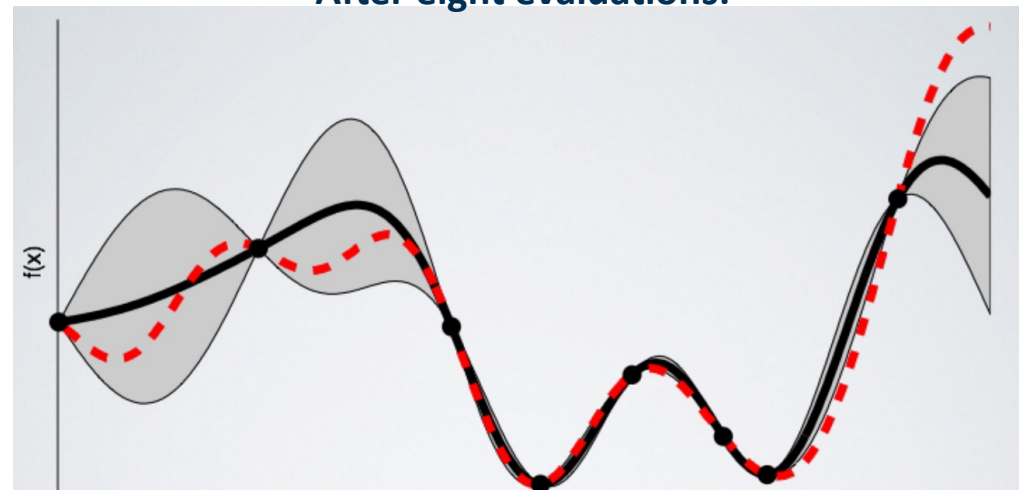
Exercise 6: Hyperparameter Tuning

04.04.2022

After two evaluations:



After eight evaluations:



Hyperparameter Tuning Task

You are training a model with 8 binary hyperparameters.

- 1) To find the optimal solution, how many runs over the complete dataset would you need with Grid Search?
How many if we assume the hyperparameters to be independent?
- 2) You have a budget of 8 runs over the complete dataset.
Describe how this would be realized with:
 - Random Search
 - Successive Halving (with the dataset as resource)
 - Bayesian Search

Genetic Algorithms

Task

You are training a model with 8 binary hyperparameters. You are currently optimizing the hyperparameters using SGA with a crossover probability of 0.5 and a mutation probability of 0 . The optimal hyperparameter configuration is $[1,0,0,0,1,1,0,1]$.

Given the following pairs of parents, how likely is it that we generate the optimal configuration as a child?

- 1) $[1,1,0,0,1,0,0,1]$ and $[1,0,0,0,1,1,1,1]$
- 2) $[0,0,0,0,1,1,1,1]$ and $[1,0,0,0,1,0,0,1]$
- 3) $[1,0,1,0,1,0,1,0]$ and $[0,0,0,0,1,1,0,1]$