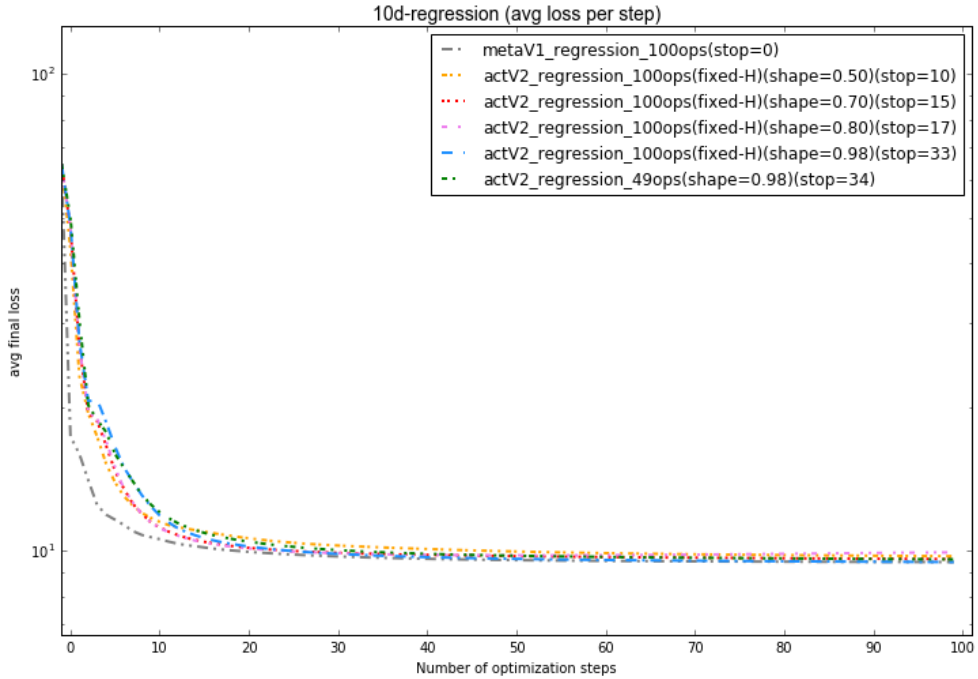
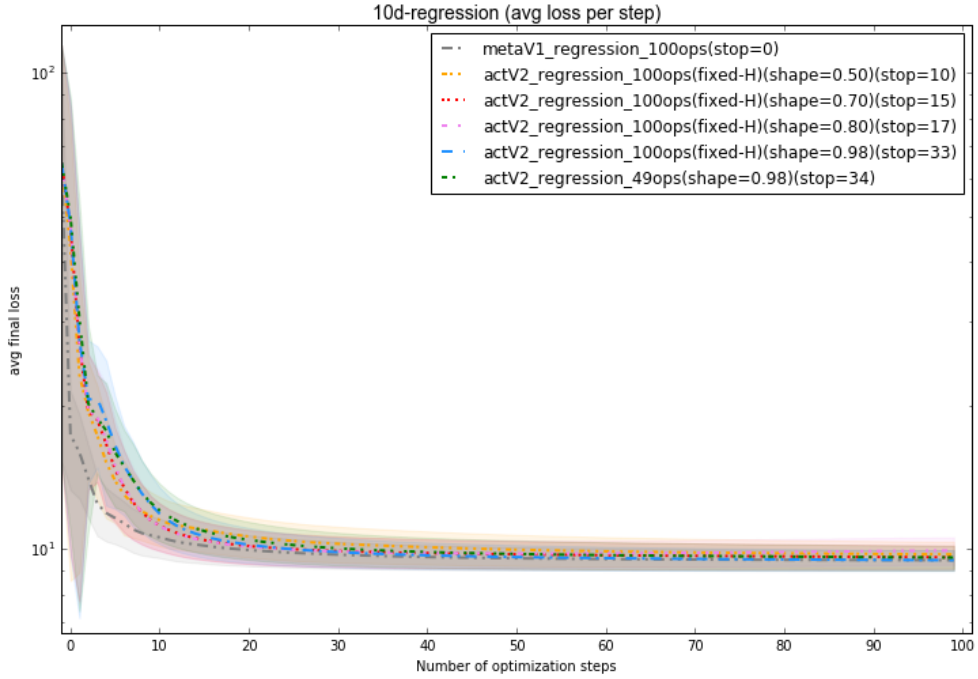
Main results of experiments with 10d-regression functions

Performance of "Meta" and "ACT" models on 20,000 freshly sampled test regression functions (right figure with +/- one standard deviation).
Models were trained for 100-150 epochs on 10,000 new sampled functions per epoch. The models were evaluated each 10th epoch on a fixed set of 20,000 validation functions. We picked the best model for the final test run.
Learning rate was set to 5e-6 (random search) and five of the six models were trained with a fixed horizon (H=100) of time steps (indicated with "fixed-H" in the legend below).
One of the ACT models was trained with a stochastic training horizon i.e. for each mini-batch we sampled a horizon H=T from a distribution p(T) with shape parameter *nu*=0.98. This resulted on average in a training horizon of 49 optimization steps (E[T]=49).

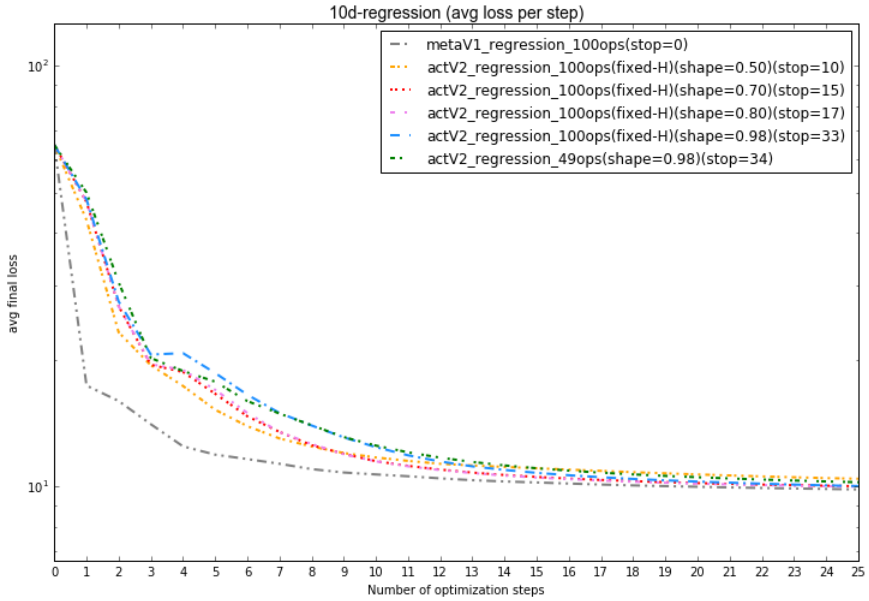
The ACT models were trained with different shape parameters for the truncated prior distribution p(t| H) which is specified in the legend with "shape".
The models used the following shape parameters {0.5, 0.7, 0.8, 0.98} (other experiments were run with 0.6, 0.85 and 0.9 but results are not shown here).

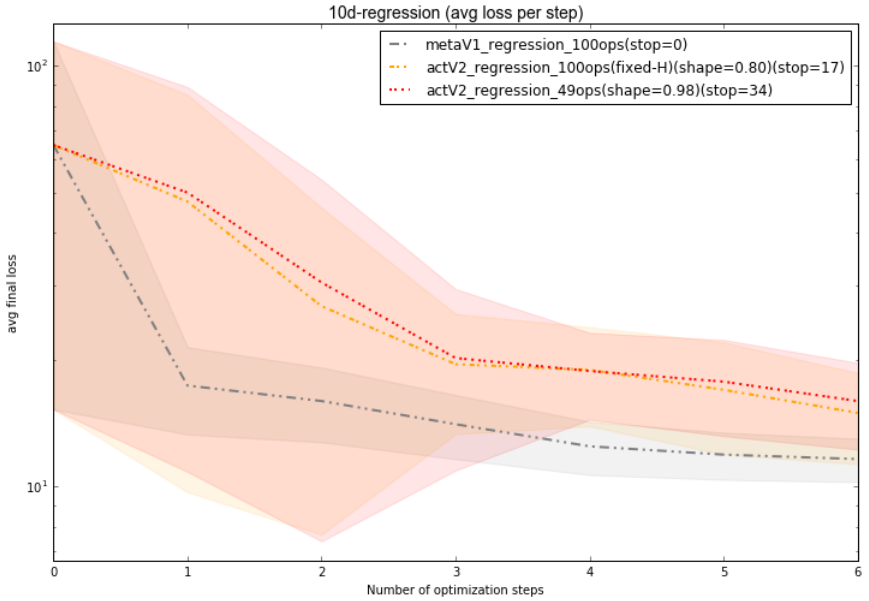




Note: that e.g. "stop=17" means that the average *stopping step* for the test functions was equal to 17 based on the cumulative PMF of the approximated q(t|H) distribution, setting the threshold to 0.95.

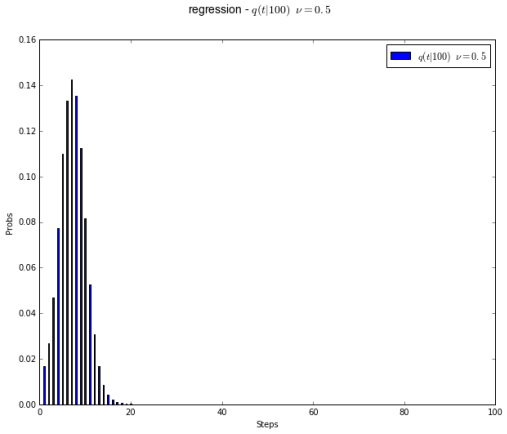
10d-regression: compare meta with ACT first optimization steps



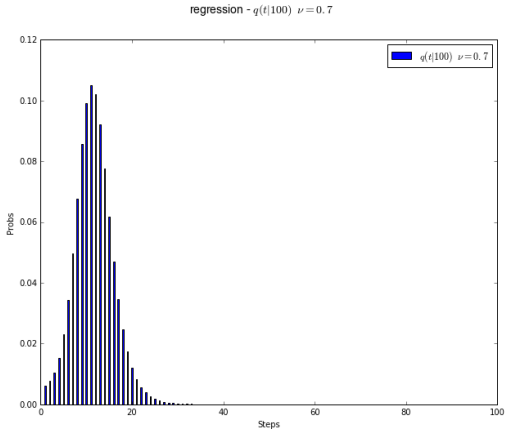


Approximated q(t|H) for ACT models on 20,000 test functions

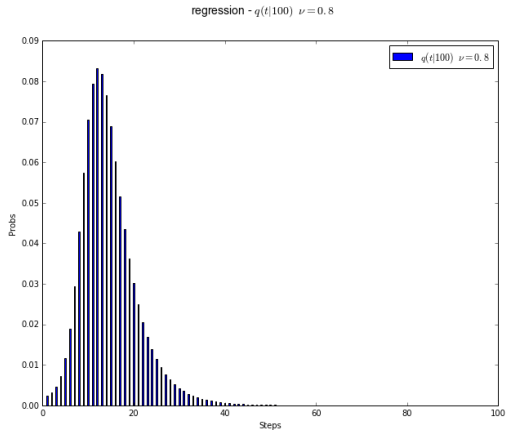
For each of the evaluated ACT models the approximated q(t | 100) distribution during the test run is shown below.
Underneath each figure we mention some statistics w.r.t. the mode of the q-distributions. **Please note** that compared to the first figure on the previous page, here two ACT models where trained with a stochastic training regime (details below).



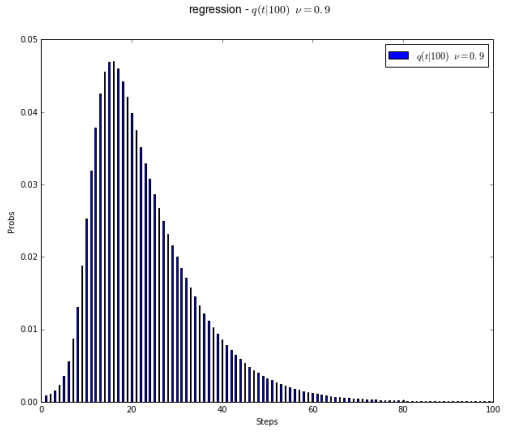
Max mode stats(N=20000): Range(1, 12) / mean=6.8 / median=7.0 / stddev=1.8



Max mode stats(N=20000): Range(3, 17) / mean=10.8 / median=11.0 / stddev=1.7

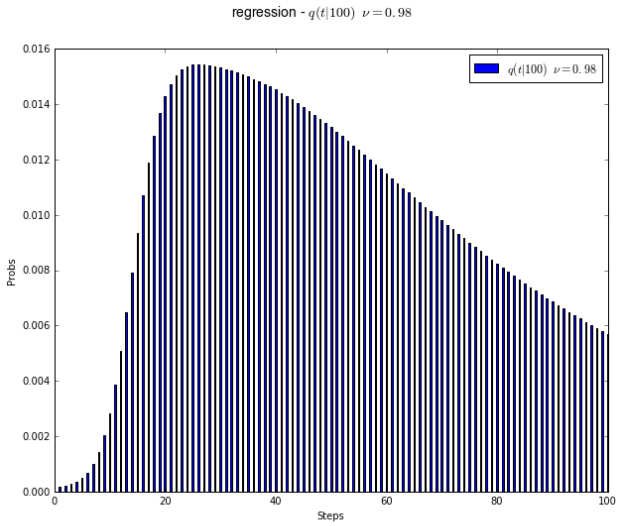


Max mode stats(N=20000): Range(4, 22) / mean=12.1 / median=12.0 / stddev=1.7

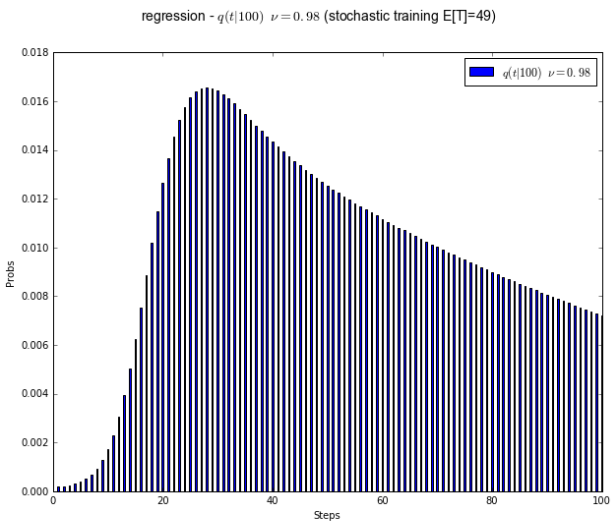


Max mode stats(N=20000): Range(9, 55) / mean=16.1 / median=15.0 / stddev=3.6

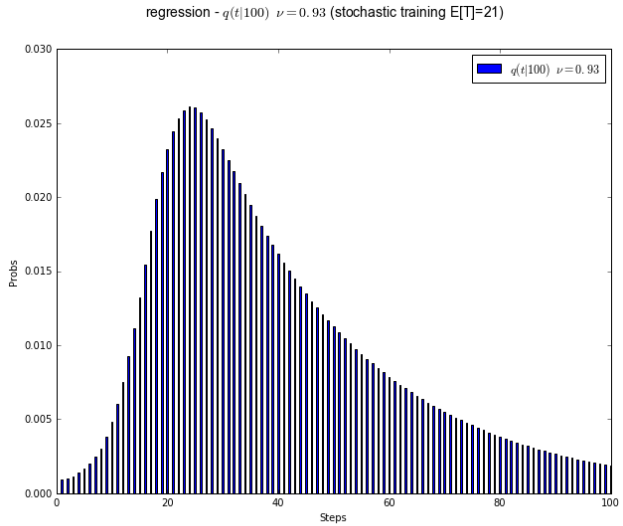
Note: Two models were trained with a stochastic training regime. For a mini-batch a horizon H was sampled from the distribution p(T) with a shape parameter *nu* equal to 0.98 and 0.93 which results in an average training horizon of E[T]=49 and E[T]=21



Max mode stats(N=20000): Range(16, 100) / mean=31.7 / median=27.0 / stddev=13.0



Max mode stats(N=20000): Range(22, 92) / mean=28.0 / median=28.0 / stddev=3.1



Max mode stats(N=20000): Range(11, 40) / mean=23.8 / median=24.0 / stddev=2.2

Detailed statistics of the approximated q(t | 100) distributions during test run on 20,000 regression functions (part 1)

For each of the ACT models trained the following plots are shown:

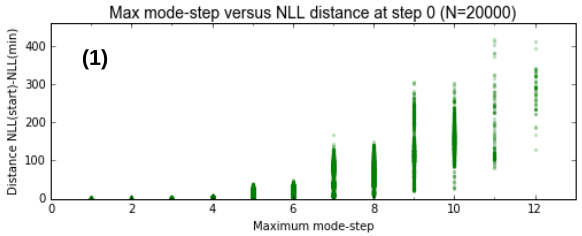
(1) Figure in which the mode-step for a function is plotted against the distance between the initial (step 0) negative log-likelihood (NLL) value and the minimum NLL value for that function.

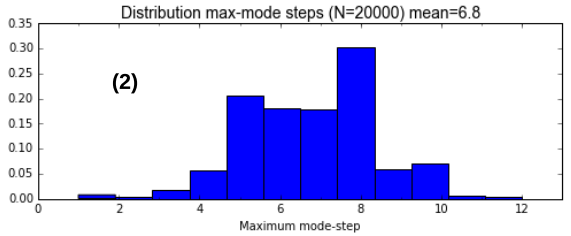
(2) Distribution of "mode steps"

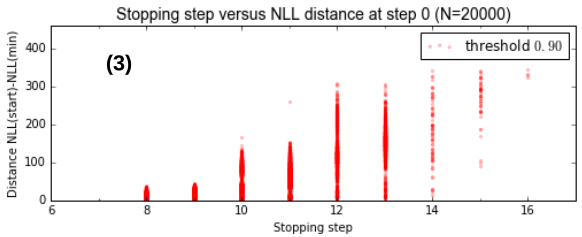
(3) Scatter plot: for each function the *stopping step* was determined based on the q(t|100) cumulative distribution using a threshold of 0.9 (as indicated in the legend). The stopping step on the x-axis was plotted against the distance between the initial NLL and the minimum NLL value for a particular function to be optimized.

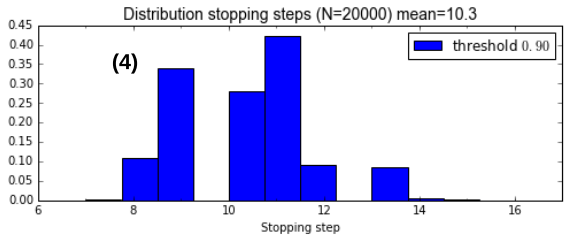
(4) Distribution of *stopping steps*

regression ($\nu = 0.5$) - 20000 functions

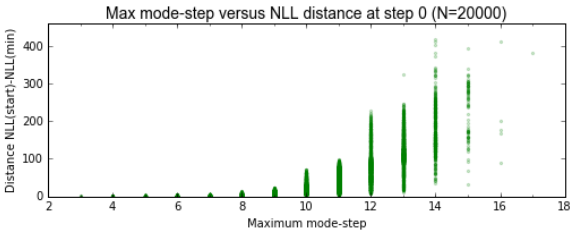


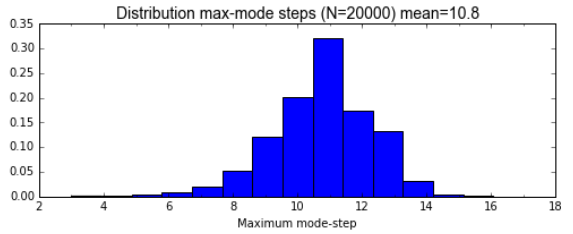


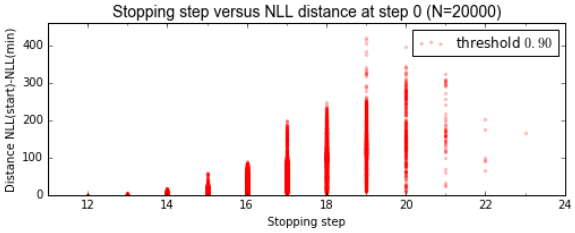


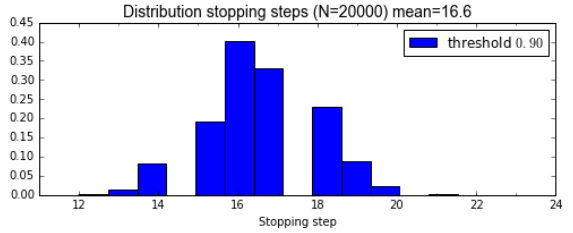


regression ($\nu = 0.7$) - 20000 functions



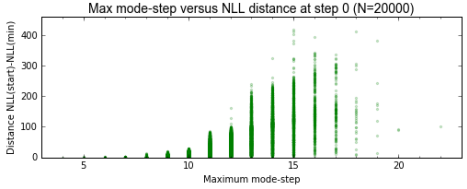


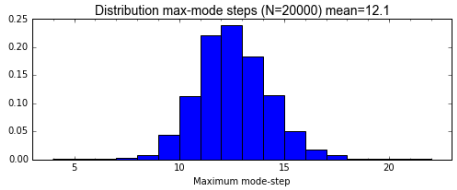


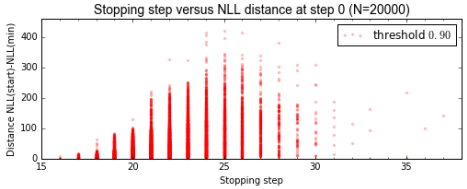


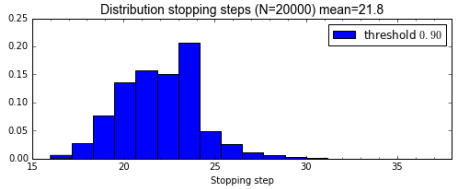
Detailed statistics of the approximated q(t | 100) distributions during test run on 20,000 regression functions (part 2)

regression ($\nu = 0.8$) - 20000 functions

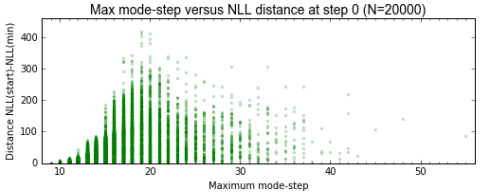


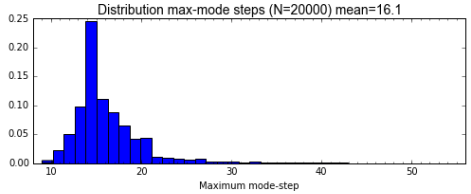


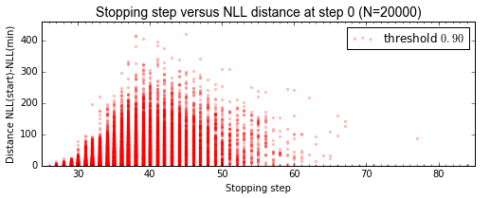


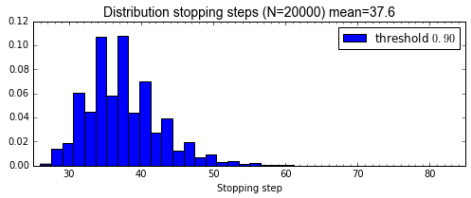


regression ($\nu = 0.9$) - 20000 functions

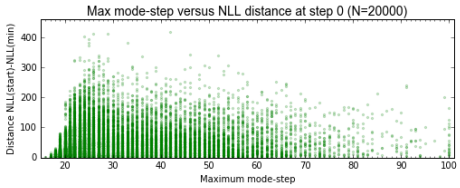


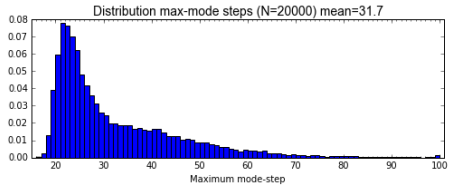


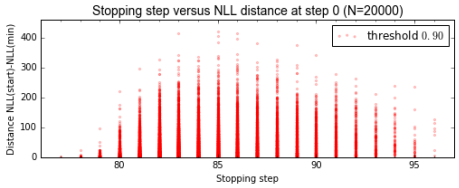


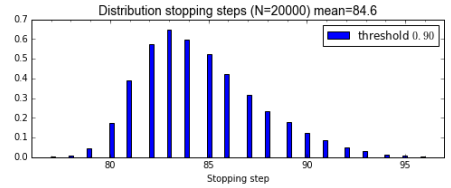


regression ($\nu = 0.98$) - 20000 functions



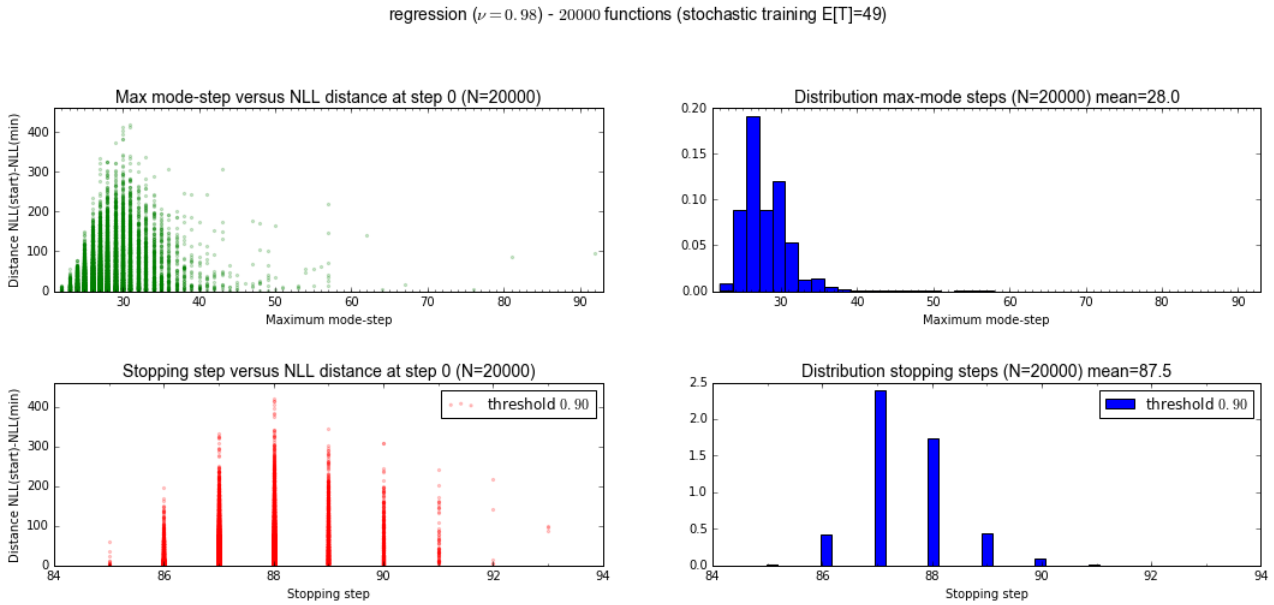




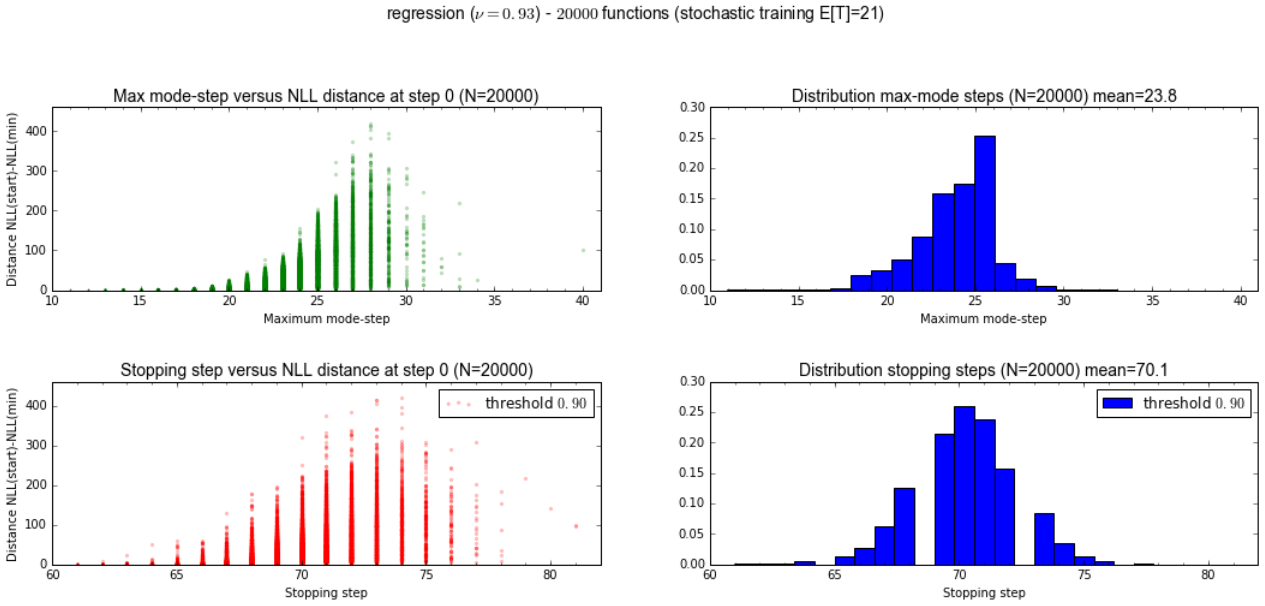


Detailed statistics of the approximated q(t | 100) distributions during test run on 20,000 regression functions (part 3)

Plots for the ACT model that was trained with a **stochastic training** regime (E[T] = 49).
Compared to the previous figure i.e. the ACT model that was trained with a fixed horizon H=100, the spread of the *mode step* and *stopping step* is significantly smaller.



Plots for the ACT model that was trained with a **stochastic training** regime (E[T] = 21).

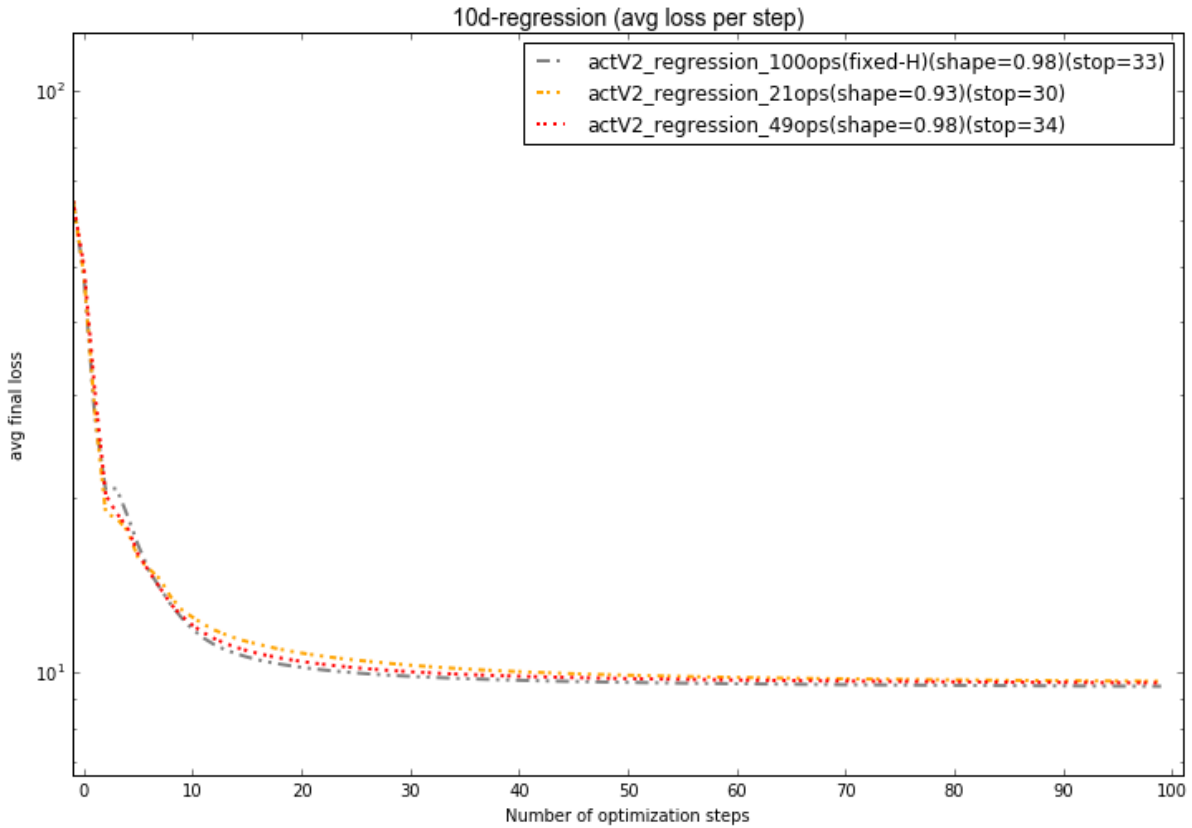


Compare performance ACT model with fixed and stochastic training horizon

Comparing the performance of three ACT models on 20,000 freshly sampled regression functions. Two models were trained with a prior p(t|H) having a shape parameter (*nu*) equal to 0.98. The other one is trained with *nu* = 0.93.

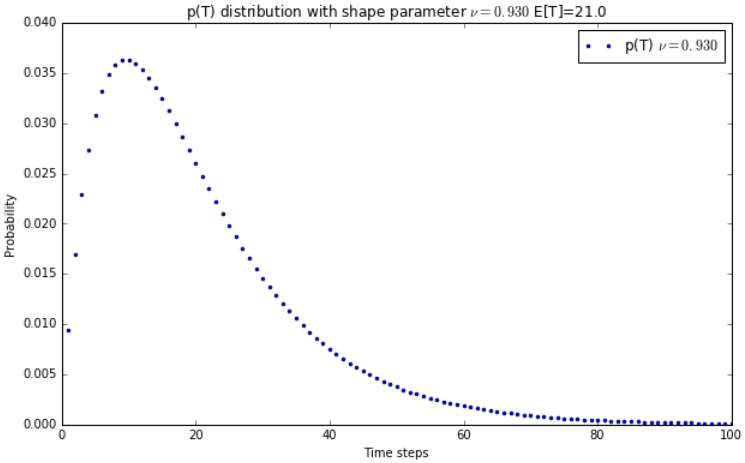
One model was trained with a fixed horizon H equal to 100. The other two where trained stochastically with E[T] = 49 and E[T]=21 optimization steps (please see legend).

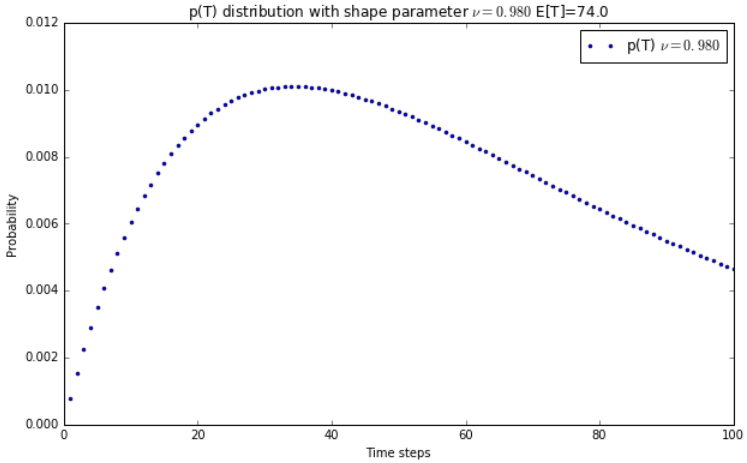
The figure on the right reveals that all models roughly perform the same although the models trained stochastically reached their optimum performance after less epochs (around 80) and obviously it took significant less computational effort to train them.



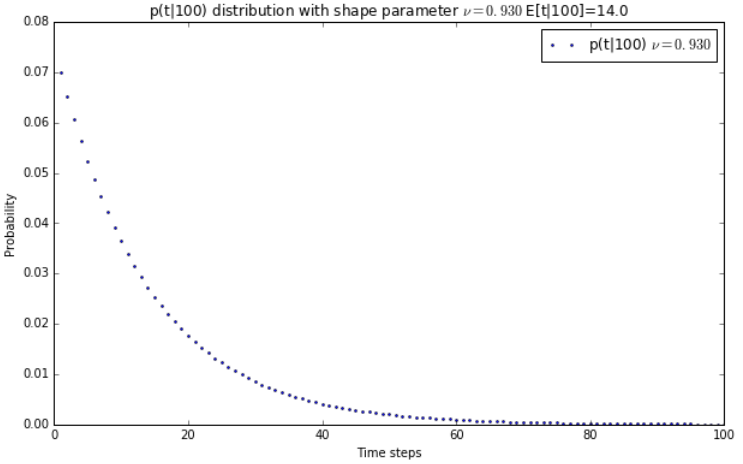
Examples of prior distributions - part 1

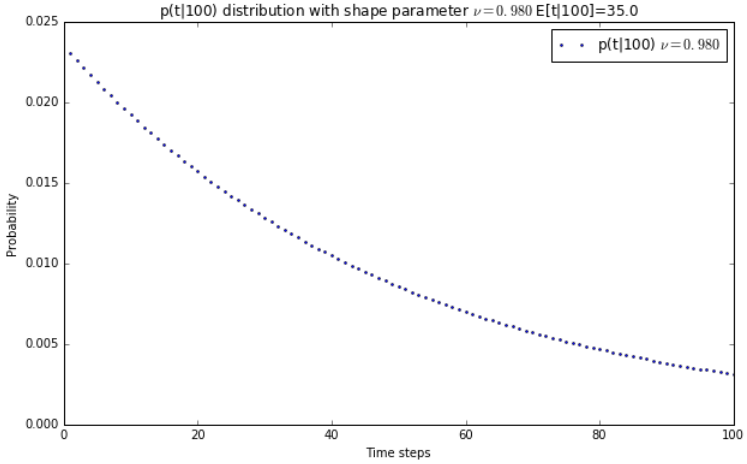
Two examples of **p(T)** distributions with different shape parameters





Two examples of **p(t|T=100)** distributions with different shape parameters





Examples of prior distributions - part 2