

Local-Fit Model Variance Estimation with Bootstrapping

Cole Weis – 3/18

Experiment:

1. Increase number of replicas (N) from 1 to some arbitrary value (5000)
2. Train a model on replica N_i
3. Record total variance of each CFF across all replicas trained so far (stdev^2)
4. Perform OLS on the function $\text{var}(N_i)/\text{var}(N_{5000})$ to produce a line of best fit describing the percentage of model variance described by N samples

Problem: Running into Rivanna time limit

- Began with:
 - One Rivanna job (slurm task) for all replicas and sets; timed out before completing

```
To execute the default application inside the container, run:  
singularity run --nv $CONTAINERDIR/tensorflow-2.1.0-py37.sif
```

```
2022-03-17 21:47:08.012378: E tensorflow/stream_executor/cuda/cuda_driver.cc:271] failed call to cuInit: CUDA_ERROR_NO_DEVICE: no CUDA-capable device is detected  
2022-03-17 21:47:08.012431: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:156] kernel driver does not appear to be running on this host (udc-ba26-35c1): /proc/driver/nvidia/versi  
2022-03-17 21:47:08.014793: I tensorflow/core/platform/cpu_feature_guard.cc:151] This TensorFlow binary is optimized with oneAPI Deep Neural Network Library (oneDNN) to use the following  
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.  
2022-03-17 21:47:09.881202: W tensorflow/python/util/util.cc:368] Sets are not currently considered sequences, but this may change in the future, so consider avoiding using them.
```

```
.....  
slurmstepd: error: *** JOB 33706785 ON udc-ba27-28c0 CANCELLED AT 2022-03-18T01:47:25 DUE TO TIME LIMIT ***
```

Problem: Running into Rivanna time limit

- Began with:
 - One Rivanna job (slurm task) for all replicas and sets; timed out before completing
- Addressing this issue:
 - Divided each replica of training into individual slurm tasks
 - Appending all data into same csv file with same format as before
- This is ~5000 slurm tasks
 - Is this too many to put onto Rivanna?
- It could also be split into 15 slurm tasks
 - Do all replicas for each set
 - Most likely would not time out but might be close

Problem: Running into Rivanna time limit

- Is the standard procedure to create a Rivanna job for each replica?
- Would the Shannon phys computer be better suited for this experiment?
 - Best for fast prototyping experiments
- `#SBATCH -t 48:00:00` increase time limit

Next Steps:

- Move experiment to BKM10 formalism
- Adapt comparison method to global fit procedures
 - Perform similar experiment with k-fold C.V.
 - Number of replicas will be substituted for number of folds
 - How many folds required to reveal a certain percent of model variance?
 - Compare bootstrapping to k-fold C.V.?
 - Number of training iterations required by each to reveal model variance
 - Ability to extrapolate models performance to new datasets (psuedodata generation)