**General structure of the dataset creation methodology**

(DISCLAIMER: The code is not tested extensively; it may still have bugs/need to be modified!)

**Goal:** Create a balanced dataset of operating points defined by PD, PG and VG classified according to their feasibility to the N-1 security constrained AC-OPF for a defined list of line contingencies

* TODO: Each test case essentially needs a careful selection of line contingencies as we want that a feasible solution exist for all time series data (otherwise it would imply that for that loading pattern the system cannot be operated securely, which is not a realistic assumption)

**Step 1:** Load time series data for loads and renewable energy and map to load buses of test case

* TODO: use different data, and different method to distribute among buses

**Step 2:** For each time step of PD (QD = fixed\_factor \* PD), solve the N-1 security constrained AC-OPF to compute feasible and optimal solution (PG\*, VG\*).

* Double checked that the solution (PD, PG\*,VG\*) is also classified as feasible to the N-1 security constrained AC-OPF using AC power flows
* Compared to AC OPF solution (no line contingencies considered); this should normally violate some constraints (otherwise the line contingencies do not change the solution)

**Step 3:** Fit a distribution to the obtained (PD,PG\*,VG\*) to draw more samples that are assessed using AC power flows

* Currently, I use a simple multivariate normal distribution to draw new samples
* TODO: if necessary, use other distributions, e.g. Gaussian mixture model; I think this will also depend on the time series data used

**Step 4:** For each of the sample, solve AC power flows and evaluate constraint violations for the N-1 security criterion

**Step 5:** Assemble dataset

* Important to remove all data for which the AC power flow has not converged

Some comments:

* Using parfor to parallelize
* Redefined tolerance for constraint violation (this is important as otherwise in Step 2 solution are falsely classified as not secure)
* Keep track of the maximum violation of constraints for each type (PG QG Vbus Sline)
* I enforce Qlimits in the AC power flow; this however leads to sometimes crashing of the power flow; therefore I have a try catch routine (which sometimes print some error messages); this is ok
* Be careful that some test cases have multiple generators connected to one bus
* Some of the outputs (PG\*, VG\*) are constant (i.e. they do not change as a function of PD); therefore in the MVD their covariance matrix entries are zero 🡪 maybe we need to consider that in NN training?