A. Appendix A

A.1. Other possible expert initializations

We investigated two additional initialization regimes for the expert network:

ScratchOP - which is a trivial way to initialize g_ϕ^{t+1} with random weights. This initialization is fast and completely clean of any bias (either good or bad) of previous data. The main drawback is that previous knowledge could help g_ϕ^{t+1} to learn better representation. There is no knowledge transfer to the new expert.

FtOP - which begin the training of the next task using the expert from the current one. That is a reasonable assumption if the distributions of consecutive tasks are similar. This initialization copies the weights of g_ϕ^t into g_ϕ^{t+1}, allowing a good initialization point for the training of the task $t+1$. This approach has two disadvantages: first, the knowledge of g_ϕ^t has a lot of recency bias (g_ϕ^t has already forgotten knowledge from tasks $t-1$); second, the architectures of g_ϕ^t and g_ϕ^{t+1} need to be homogenous to perform the copy operation.

Table 5 show results of ScratchOP and FtOP expert initialization for POCON in CIFAR-100. The results for CaSSLe and PFR are recall once again here for an easy comparison.

Table 5: Accuracy of a linear evaluation on split CIFAR-100 for different number of tasks.

CIFAR-100 (32x32)					
Method	**4 tasks**	**10 tasks**	**20 tasks**	**50 tasks**	**100 tasks**
FT	54.8	50.94	44.95	38.0	27.0
CaSSLe	59.80	52.5	49.6	45.3	42.10
PFR	59.70	54.33	44.80	46.5	43.30
POCON					
ScratchOP	57.64	46.87	39.69	30.42	30.50
FtOP	62.0	58.13	55.57	48.64	47.06
Joint	65.4				

A.2. Data partition for task-free setting experiment

Fig. 4 depicts a data partition for 10 tasks and beta 4 based on [63]. The y-axis denotes the probability of selecting a sample in an iteration. Hence, the batches of data will contain a mixture of tasks data proportional to the probability of taking a sample for each class (beta= 4 produces a mixture of two tasks). Fig. 4 shows where the mixed batches are taken, for instance at 8000 iteration for this particular instance. In such created setting, there are no explicit tasks' boundaries that could be used in the continual training session to initiate additonal step for training a new task.

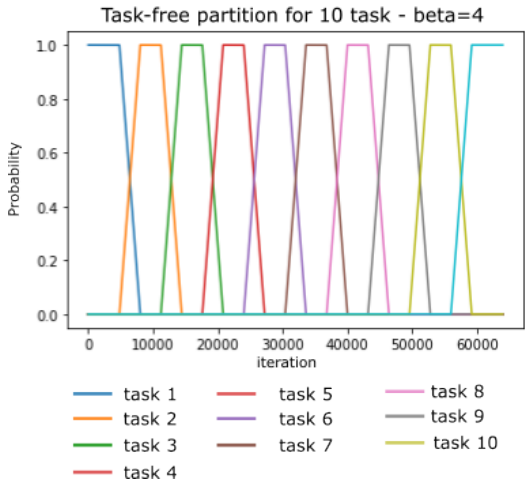

Figure 4: Sample data partition for 10 task and beta 4