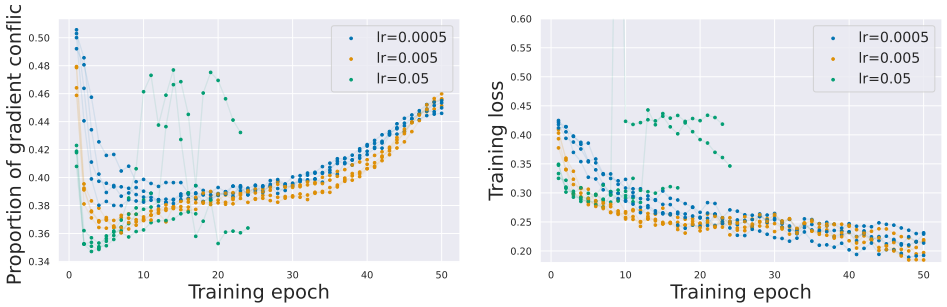
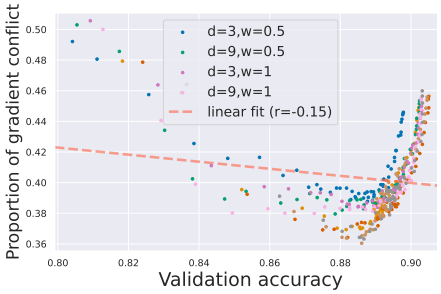
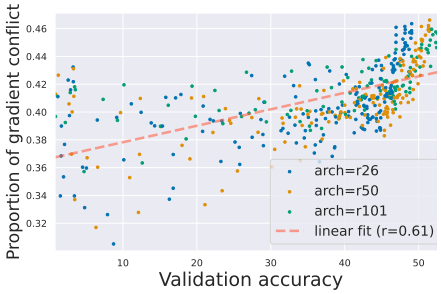


(a) Gradient conflicts *(left)* and training loss *(right)* for uniform MTL on the 40 tasks of CelebA for different learning rates (colors; $lr \in [5e^{-4}, 5e^{-3}, 5e^{-2}]$) and model sizes (lines; depth $\in [3, 9]$, width $\in [0.5, 1]$).



(b) Proportion of gradient conflicts as a function of accuracy, averaged across CelebA's 40 tasks.



(c) Proportion of gradient conflicts as a function of accuracy, averaged across for DomainNet's 6 domains.

Figure 1: Additional measurements of the proportion of gradient conflicting pairs of tasks/domains, with increased number of tasks and varying model sizes and learning rates. Gradient conflict is measured following the definition of PCGrad (*Gradient Surgery*) as implemented in `https://github.com/VICO-UoE/UniversalRepresentations`

Table 1: Impact of the population size N in PBT on the result of the scalarization weights search (*left*) and on its computational cost, when compared to multi-task optimization methods (*right*).

	N=6	N=12	N=24	N=40
E=3, Q=0.25	90.320	**90.375**	90.345	90.370
E=5, Q=0.25	90.315	90.355	90.330	**90.380**
average	90.317	90.365	90.337	**90.375**

(a) Average test accuracy (2 seeds) when training on all **40** CelebA tasks with the scalarization weights policy found by PBT for different N and E E. For reference, the corresponding uniform MTL baseline yields **90.303** accuracy, and the random loss weighing baseline (RLW) **90.327**, both averaged across 4 random seeds.

Costs	# forward passes	# backward passes	additional computations	additional storage
Uniform MTL	1	1	0	1
Uncertainty	1	1	minor	1
PBT with N models	$N+1$	$N+1$	minor (checkpoint writing)	1
PCGrad	1	T	T^2	T
RotoGrad	1	T	T	T
GradDrop	1	T	0	T

(b) Theoretical costs per training iteration relative to the vanilla uniform MTL baseline; expressed in terms of compute (number of forward passes, backward passes, and additional computations such as computing gradients conflicts) and memory (e.g. storing per-task gradients simultaneously), as a function of the number of tasks T.

Table 2: Comparing the outcome of PBT ($E = 3, Q = 0.25$) and grid search when training a MTL model on **three** tasks/attributes of CelebA with quantitative results on the *left* (results are averaged across 3 random seeds) and an illustration of the search space covered by PBT on the *right*

Method	PBT ($N = 6$)	PBT ($N = 24$)	Grid ($N = 8$)	Grid ($N = 125$)
Accuracy	84.31	84.45	84.29	84.50