

## A APPENDIX

### A.1 INITIATOR PBT ON ROSENBROCK

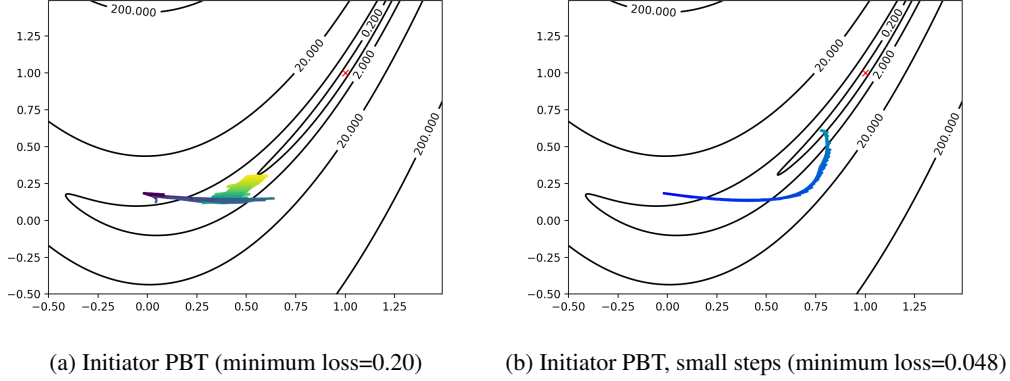


Figure 2: Trajectories of 100 PBT training steps (16 jobs per step) on the Rosenbrock function with  $a = 1$  and  $b = 100$  (minimum at the red cross  $(1, 1)$ , trajectories go from blue to green)

### A.2 LANGUAGE MODELING ON PENN TREE BANK

Table 3: The dropouts from Transformer-XL that we tune through PBT.

dropouta	applied to multi-head attention layers
dropoute	to remove words from embedding layer
dropoutff	applied to positionwise ff layers
dropouti	for input embedding vectors
dropouto	applied to the output (before the logit)

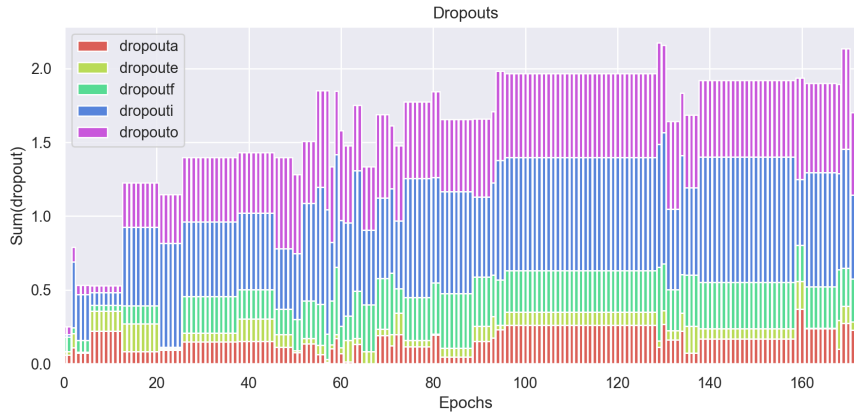


Figure 3: Dropout schedule of the best run of ROMUL 32 workers on PTB

### A.3 REGULARIZATION SCHEDULES ON WIKITEXT-103

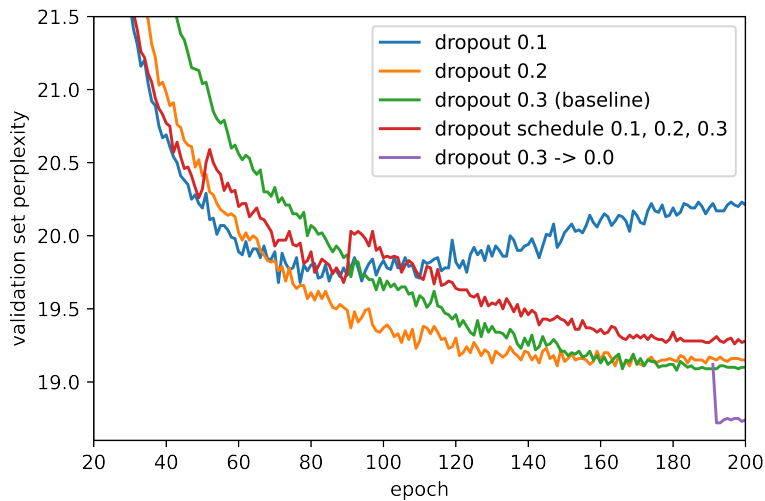


Figure 4: Lower dropout values are better early, but are outperformed by more strongly regularized models later (red, orange and blue lines) - here on wikitext103 with a 247M parameters language model from Fan et al. (2019) (*Adaptive Inputs + LayerDrop*). PBT algorithms would tend to reduce dropout aggressively early on: after that, even if the dropout is increased later, the performance remains worse than training with a high dropout from the beginning (red line). Perhaps counterintuitively, this hints against increasing regularization over the course of the training - in the opposite, we observe that fine-tuning the model without dropout significantly improves test performance (purple line reaches 17.98 test perplexity) compared to the baseline (green: 18.42 test perplexity)