Algorithm: Lightweight data collection for MAP regression (interleaved pruning).

Input: Pre-trained model Y_0, number of rounds rds, pruning budget k, total layers L

Output: Training dataset $train\_data$

Initialize $train\_data \leftarrow \{(\tilde{m}_0^a, \tilde{m}_0^g, a_0)\}$;

Function `PruneIterative`$((\tilde{m}_0^a, \tilde{m}_0^g),\ (\tilde{m}_{\max}^a, \tilde{m}_{\max}^g))$**:**

  for $n \leftarrow 1$ **to** rds **do**

    `// Step 1: prune attention`

    $\tilde{m}_n^a \leftarrow \tilde{m}_0^a + n \cdot \frac{\tilde{m}_{\max}^a - \tilde{m}_0^a}{rds}$;

    $\tilde{m}_{n-1}^g \leftarrow \tilde{m}_0^g + (n-1) \cdot \frac{\tilde{m}_{\max}^g - \tilde{m}_0^g}{rds}$;

    Prune Y_{n-1} along *attention* to ratio $\tilde{m}_n^a$ to obtain Y_n';

    `// Fine-tuning and evaluation`

    Fine-tune Y_n' and evaluate $\rightarrow (\tilde{m}_n^a, \tilde{m}_{n-1}^g, a_n)$;

    Append $(\tilde{m}_n^a, \tilde{m}_{n-1}^g, a_n)$ to $train\_data$;

    `// Step 2: prune activation`

    $\tilde{m}_n^g \leftarrow \tilde{m}_0^g + n \cdot \frac{\tilde{m}_{\max}^g - \tilde{m}_0^g}{rds}$;

    Prune Y_n' along *activation* to ratio $\tilde{m}_n^g$ to obtain Y_n;

    `// Fine-tuning and evaluation`

    Fine-tune Y_n and evaluate $\rightarrow (\tilde{m}_n^a, \tilde{m}_n^g, a_n)$;

    Append $(\tilde{m}_n^a, \tilde{m}_n^g, a_n)$ to $train\_data$;

  end

Set $\tilde{m}_{a,\max} \leftarrow k/L$ and $\tilde{m}_{g,\max} \leftarrow k/L$;

`PruneIterative`$((\tilde{m}_0^a, \tilde{m}_0^g),\ (\tilde{m}_{a,\max}, \tilde{m}_{g,\max}))$;

return $train\_data$;