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

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

Initialize $train\_data = \{(\tilde{m}_{a,0}, \tilde{m}_{g,0}, a_0)\}$;

Function `PruneIterative`$((\tilde{m}_{a,0}, \tilde{m}_{g,0}),\ (\tilde{m}_{a,\max}, \tilde{m}_{g,\max}))$**:**

  for $n = 1$ **to** rds **do**

    `// Step 1: prune attention`

    $\tilde{m}_{a,n} = \tilde{m}_{a,0} + n \cdot \frac{\tilde{m}_{a,\max} - \tilde{m}_{a,0}}{rds}$;

    $\tilde{m}_{g,n-1} = \tilde{m}_{g,0} + (n-1) \cdot \frac{\tilde{m}_{g,\max} - \tilde{m}_{g,0}}{rds}$;

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

    `// Fine-tuning and evaluation`

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

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

    `// Step 2: prune activation`

    $\tilde{m}_{g,n} = \tilde{m}_{g,0} + n \cdot \frac{\tilde{m}_{g,\max} - \tilde{m}_{g,0}}{rds}$;

    Prune Y'_n along *activation* to ratio $\tilde{m}_{g,n}$ to obtain Y_n;

    `// Fine-tuning and evaluation`

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

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

  end

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

`PruneIterative`$((\tilde{m}_{a,0}, \tilde{m}_{g,0}),\ (\tilde{m}_{a,\max}, \tilde{m}_{g,\max}))$;

return $train\_data$