DTO-KD: DYNAMIC TRADE-OFF OPTIMIZATION FOR EFFECTIVE KNOWLEDGE DISTILLATION SUPPLEMENTARY MATERIAL

Zeeshan Hayder[1,2]**, Ali Cheraghian**[1,2]**, Lars Petersson**[1]**, Mehrtash Harandi**[3]**, Richard Hartley**[2]
[1]Data61 / CSIRO, Australia
[2]Australian National University (ANU), Australia
[3]Monash University, Australia

1 APPENDIX

In this section, we provide a more in-depth theoretical analysis of the proposed DTO-KD optimisation approach.

THEORETICAL ANALYSIS OF THE CLOSED FORM FORMULATION

In this section, we provide a detailed formulation for the closed-form solution.

Theorem 1.1 (Closed Form Solution)**.** *Let* $\boldsymbol{J}_t = [\nabla \log\big(L_{\mathit{distill}}(\boldsymbol{\theta}_t)\big), \nabla \log\big(L_{\mathit{task}}(\boldsymbol{\theta}_t)\big)] \in \mathbb{R}^{n\times 2}$. *The closed-form solution to the optimization problem*

$$\begin{aligned} \boldsymbol{\pi}^* \in \arg\min_{\boldsymbol{\pi}} \frac{1}{2}\big\|\boldsymbol{J}_t\boldsymbol{\pi}\big\|^2 \\ \textit{s.t.} \quad \pi_1 + \pi_2 = 1 \end{aligned} \tag{1}$$

is given by

$$\pi_1^* = \frac{g_{22} - g_{12}}{g_{11} + g_{22} - 2g_{12}}\ , \tag{2}$$

$$\pi_2^* = \frac{g_{11} - g_{12}}{g_{11} + g_{22} - 2g_{12}}\ , \tag{3}$$

where $\boldsymbol{G} = \boldsymbol{J}_t^\top \boldsymbol{J}_t$ *is the Gram matrix:*

$$\boldsymbol{G} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix},$$

with elements

$$\begin{aligned} g_{11} &= \big\|\nabla \log\big(L_{\mathit{distill}}(\boldsymbol{\theta}_t)\big)\big\|^2, \\ g_{12} = g_{21} &= \big\langle \nabla \log\big(L_{\mathit{distill}}(\boldsymbol{\theta}_t)\big), \nabla \log\big(L_{\mathit{task}}(\boldsymbol{\theta}_t)\big)\big\rangle, \\ g_{22} &= \big\|\nabla \log\big(L_{\mathit{task}}(\boldsymbol{\theta}_t)\big)\big\|^2. \end{aligned}$$

Proof. Using the constraint $\pi_2 = 1 - \pi_1$, we can expand the objective function as follows:

$$\begin{aligned} \frac{1}{2}\left\|\boldsymbol{J}_t\boldsymbol{\pi}\right\|^2 &= \frac{1}{2}\left\|\pi_1 \nabla \log\big(\mathrm{L}_{\mathrm{distill}}(\boldsymbol{\theta}_t)\big) + \pi_2 \nabla \log\big(\mathrm{L}_{\mathrm{task}}(\boldsymbol{\theta}_t)\big)\right\|^2 \\ &= \frac{1}{2}\left(\pi_1^2 g_{11} + \pi_2^2 g_{22} + 2\pi_1\pi_2 g_{12}\right) \\ &= \frac{1}{2}\left(\pi_1^2 g_{11} + (1-\pi_1)^2 g_{22} + 2\pi_1(1-\pi_1) g_{12}\right) \end{aligned}$$

Taking the derivative with respect to π_1 and setting it to zero:

$$\begin{aligned}\frac{d}{d\pi_1}\frac{1}{2}\left(\pi_1^2 g_{11} + (1-\pi_1)^2 g_{22} + 2\pi_1(1-\pi_1)g_{12}\right) = \\ \pi_1\left(g_{11}+g_{22}-2g_{12}\right) - g_{22} + g_{12} \\ \implies \boxed{\pi_1^* = \frac{g_{22}-g_{12}}{g_{11}+g_{22}-2g_{12}}}\ .\end{aligned}$$

Using the fact that $\pi_1 + \pi_2 = 1$, we obtain

$$\pi_2^* = \frac{g_{11}-g_{12}}{g_{11}+g_{22}-2g_{12}}\ .$$

□

Corollary 1.2 (Alignment of $\boldsymbol{g}^*$)**.** *Define $\boldsymbol{g}_1 = \nabla \log\left(L_{distill}(\boldsymbol{\theta}_t)\right)$ and $\boldsymbol{g}_2 = \nabla \log\left(L_{task}(\boldsymbol{\theta}_t)\right)$. Then the update direction $\boldsymbol{g}^* = \pi_1\boldsymbol{g}_1 + \pi_2\boldsymbol{g}_2$ for $\boldsymbol{\pi}^*$ defined in Equation (2) is aligned with both $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$.*

Proof. Assume that $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ are in conflict (*i.e.*, $-1 < \langle \boldsymbol{g}_1, \boldsymbol{g}_2\rangle < 0$). We study the behavior of $\langle \boldsymbol{g}^*, \boldsymbol{g}_1\rangle 0$ and $\langle \boldsymbol{g}^*, \boldsymbol{g}_2\rangle$ below.

$$\begin{aligned}\langle \boldsymbol{g}^*, \boldsymbol{g}_1\rangle &= \langle \pi_1\boldsymbol{g}_1 + \pi_2\boldsymbol{g}_2, \boldsymbol{g}_1\rangle = \pi_1\|\boldsymbol{g}_1\|^2 + \pi_2\langle \boldsymbol{g}_2, \boldsymbol{g}_1\rangle \\ &= \frac{g_{22}-g_{12}}{g_{11}+g_{22}-2g_{12}}g_{11} + \frac{g_{11}-g_{12}}{g_{11}+g_{22}-2g_{12}}g_{12} \\ &= \frac{g_{11}g_{22}-g_{12}^2}{\|\boldsymbol{g}_1-\boldsymbol{g}_2\|^2}\end{aligned}$$

Similarly, for $\boldsymbol{g}_2$:

$$\begin{aligned}\langle \boldsymbol{g}^*, \boldsymbol{g}_2\rangle &= \langle \pi_1\boldsymbol{g}_1 + \pi_2\boldsymbol{g}_2, \boldsymbol{g}_2\rangle = \pi_1\langle \boldsymbol{g}_1, \boldsymbol{g}_2\rangle + \pi_2\|\boldsymbol{g}_2\|^2 \\ &= \frac{g_{22}-g_{12}}{g_{11}+g_{22}-2g_{12}}g_{12} + \frac{g_{11}-g_{12}}{g_{11}+g_{22}-2g_{12}}g_{22} \\ &= \frac{g_{11}g_{22}-g_{12}^2}{\|\boldsymbol{g}_1-\boldsymbol{g}_2\|^2}.\end{aligned}$$

Note that $g_{11}g_{22} - g_{12}^2 = \det(\boldsymbol{J}^\top\boldsymbol{J})$ and since $\boldsymbol{G} = \boldsymbol{J}^\top\boldsymbol{J}$ is always positive semi-definite, then $\det(\boldsymbol{J}^\top\boldsymbol{J}) \geq 0$. This proves that $\boldsymbol{g}^*$ is always positively aligned with both $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$. □

Corollary 1.3 (Equal Contribution of $\boldsymbol{g}^*$ to Both Losses)**.** *In Corollary 1.2, we showed that*

$$\langle \boldsymbol{g}^*, \boldsymbol{g}_1\rangle = \langle \boldsymbol{g}^*, \boldsymbol{g}_2\rangle = \frac{g_{11}g_{22}-g_{12}^2}{\|\boldsymbol{g}_1-\boldsymbol{g}_2\|^2}.$$

This implies that the update direction contributes equally to the descent of both the distillation and task losses, effectively mitigating gradient dominance.

Corollary 1.4 (Lower Bound on $\|\boldsymbol{g}^*\|$)**.** *The norm of the optimal update direction $\boldsymbol{g}^*$ satisfies the lower bound:*

$$\|\boldsymbol{g}^*\| \geq \frac{1}{\sqrt{2}}\min(\|\boldsymbol{g}_1\|, \|\boldsymbol{g}_2\|)\ . \tag{4}$$

This implies that the update magnitude remains controlled and does not collapse under gradient imbalance.

Proof. We have:

$$\|\boldsymbol{g}^*\|^2 = \frac{\|\boldsymbol{g}_1\|^2\|\boldsymbol{g}_2\|^2}{\|\boldsymbol{g}_1-\boldsymbol{g}_2\|^2}\ .$$

Note that $\|\boldsymbol{g}_1 - \boldsymbol{g}_2\|^2 \leq \|\boldsymbol{g}_1\|^2 + \|\boldsymbol{g}_2\|^2$ and hence

$$\|\boldsymbol{g}^*\|^2 \geq \frac{\|\boldsymbol{g}_1\|^2\|\boldsymbol{g}_2\|^2}{\|\boldsymbol{g}_1\|^2 + \|\boldsymbol{g}_2\|^2} \ .$$

Without loss of generality, assume $\|\boldsymbol{g}_1\| \leq \|\boldsymbol{g}_2\|$. We have

$$\begin{aligned}\|\boldsymbol{g}^*\|^2 &\geq \frac{\|\boldsymbol{g}_1\|^2\|\boldsymbol{g}_2\|^2}{\|\boldsymbol{g}_1\|^2 + \|\boldsymbol{g}_2\|^2} \\ &\geq \frac{\|\boldsymbol{g}_1\|^2\|\boldsymbol{g}_2\|^2}{2\max\left(\|\boldsymbol{g}_1\|^2, \|\boldsymbol{g}_2\|^2\right)} = \frac{1}{2}\min\left(\|\boldsymbol{g}_1\|^2, \|\boldsymbol{g}_2\|^2\right)\end{aligned}$$

As such,

$$\|\boldsymbol{g}^*\| \geq \frac{1}{\sqrt{2}}\min(\|\boldsymbol{g}_1\|, \|\boldsymbol{g}_2\|) \ .$$

Thus, the update direction always maintains a lower bound, ensuring that it does not collapse even when one gradient is small. □

Corollary 1.5 (Upper Bound on $\|\boldsymbol{g}^*\|$)**.** *The norm of the optimal update direction $\boldsymbol{g}^*$ satisfies the upper bound:*

$$\|\boldsymbol{g}^*\| \leq \frac{\|\boldsymbol{g}_1\|\|\boldsymbol{g}_2\|}{\left|\|\boldsymbol{g}_1\| - \|\boldsymbol{g}_2\|\right|}. \tag{5}$$

As such, the magnitude of the updates does not grow excessively when the gradients have different scales.

Proof. Note that

$$\begin{aligned}\|\boldsymbol{g}_1 - \boldsymbol{g}_2\|^2 &= \|\boldsymbol{g}_1\|^2 + \|\boldsymbol{g}_2\|^2 \\ -2 \underbrace{\langle \boldsymbol{g}_1, \boldsymbol{g}_2 \rangle}_{\leq \|\boldsymbol{g}_1\|\|\boldsymbol{g}_2\|} &\geq (\|\boldsymbol{g}_1\| - \|\boldsymbol{g}_2\|)^2.\end{aligned}$$

Hence

$$\begin{aligned}\|\boldsymbol{g}^*\|^2 = \frac{\|\boldsymbol{g}_1\|^2\|\boldsymbol{g}_2\|^2}{\|\boldsymbol{g}_1 - \boldsymbol{g}_2\|^2} &\leq \frac{\|\boldsymbol{g}_1\|^2\|\boldsymbol{g}_2\|^2}{(\|\boldsymbol{g}_1\| - \|\boldsymbol{g}_2\|)^2} \\ &\implies \boxed{\|\boldsymbol{g}^*\| \leq \frac{\|\boldsymbol{g}_1\|\|\boldsymbol{g}_2\|}{\left|\|\boldsymbol{g}_1\| - \|\boldsymbol{g}_2\|\right|}} \ .\end{aligned}$$

□