AdaDARE-γ: Balancing Stability and Plasticity in Multi-modal LLMs through Efficient Adaptation

Supplementary Material

A. Proof

A.1. Proof of Theorem 1

Theorem 1 Consider the optimization problem of minimizing the expected layer-wise loss:

$$\arg\min_{P^\ell} \mathbb{E}\left[\left\|\Theta^\ell_{\text{sft}}\mathcal{X}^\ell_{\mathcal{T}} - \Theta^\ell_{\text{fusion}}\mathcal{X}^\ell_{\mathcal{T}}\right\|_2^2\right], \tag{A.1}$$

subject to the constraints:

$$\frac{\sum_{i=1}^n p_i}{n} >= p, \quad \text{and} \quad 0 \le p_i < 1 \quad \forall i. \tag{A.2}$$

where p represents the desired sparsity ratio. Then, the optimal probabilities p_i^* that minimize the expected loss are given by:

$$p_i^* = \max(0, 1 - \frac{n(1-p)\sqrt{H_{ii}\delta_i^2}}{\sum_{j=1}^n \sqrt{H_{jj}\delta_j^2}}) \quad \forall i. \tag{A.3}$$

Here, H_{ii} represents the i-th diagonal element of the Hessian matrix, and δ_i is the i-th element of $\Delta\Theta^{\mathcal{T},\ell}$.

Proof. First, we derive a simpler form of the loss function

$$\arg\min_{P^\ell} \mathbb{E}\left[\left\|\Theta^\ell_{\text{sft}}\mathcal{X}^\ell_{\mathcal{T}} - \Theta^\ell_{\text{fusion}}\mathcal{X}^\ell_{\mathcal{T}}\right\|_2^2\right], \tag{A.4}$$

By leveraging that the gradient term at Θ^l_{sft} equals zero, and $\mathbb{E}\left[\Theta^l_{\text{sft},i} - \Theta^l_{\text{fusion},i}\right] = (1-\gamma)\,\Delta\Theta_i^{\mathcal{T},\ell}$ is independent of p_i, we obtain:

$$\arg\min_{p_i} \mathbb{E}\left[\sum_{i=1}^n H_{ii}\left(\Theta^l_{\text{sft},i} - \Theta^l_{\text{fusion},i}\right)^2\right], \tag{A.5}$$

To minimize Eq. (A.5) under the given constraints, we then compute the expected squared difference:

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}\left[\sum_{i=1}^n H_{ii}\left(\Theta^l_{\text{sft},i} - \Theta^l_{\text{fusion},i}\right)^2\right] \\
&= \sum_{i=1}^n H_{ii}\left[(1-p_i)\delta_i^2\left(\frac{\gamma-(1-p_i)}{1-p_i}\right)^2 + p_i\delta_i^2\right] \\
&= \sum_{i=1}^n H_{ii}\delta_i^2(\frac{\alpha^2+2\alpha p_i+p_i^2}{1-p_i} + p_i)
\end{aligned} \tag{A.6}$$

where we denote $\alpha = \gamma - 1$ for simplicity. To solve this constrained optimization problem, we construct the Lagrangian:

$$\begin{aligned}
\mathcal{L} = \sum_{i=1}^n H_{ii}\delta_i^2\left(\frac{\alpha^2+2\alpha p_i+p_i^2}{1-p_i} + p_i\right) + \\
\lambda\left(p - \frac{1}{n}\sum_{i=1}^n p_i\right) - \sum_{i=1}^n \mu_i p_i,
\end{aligned} \tag{A.7}$$

where $\lambda \ge 0$, $\mu_i \ge 0$ are the Lagrange multipliers for the sparsity, non-negativity constraint respectively. By deriving the KKT conditions:

Stationarity Condition for each i:

$$H_{ii}\delta_i^2\phi'(p_i) - \frac{\lambda}{n} - \mu_i = 0 \tag{A.8}$$

where:

$$\begin{aligned}
\phi'(p_i) &= \frac{d}{dp_i}\left[\frac{(\alpha+p_i)^2}{1-p_i} + p_i\right] \\
&= \frac{\left[2(\alpha+p_i)(1-p_i) + (\alpha+p_i)^2\right]}{(1-p_i)^2} \\
&= \frac{(\alpha+1)^2}{(1-p_i)^2} \\
&= \frac{\gamma^2}{(1-p_i)^2}
\end{aligned} \tag{A.9}$$

Complementary Slackness:

$$\mu_i p_i = 0. \tag{A.10}$$

Primal and Dual Feasibility:

$$0 \le p_i < 1, \lambda \ge 0, \mu_i \ge 0. \tag{A.11}$$

For the case where $p_i > 0$ (thus $\mu_i = 0$), we have:

$$\frac{H_{ii}\delta_i^2\gamma^2}{1-p_i^2} = \frac{\lambda}{n} \tag{A.12}$$

This yields:

$$p_i = 1 - \sqrt{\frac{nH_{ii}\delta_i^2\gamma^2}{\lambda}} \tag{A.13}$$

Applying the sparsity constraint:

$$\sum_{i=1}^n (1-p_i) = n(1-p). \tag{A.14}$$

and substituting $1-p_i$:

$$\sum_{i=0}^{n} 1-\sqrt{\frac{nH_{ii}\delta_i^2\gamma^2}{\lambda}} = n\left(1-p\right). \tag{A.15}$$

We can solve for λ:

$$\begin{aligned} \sqrt{\lambda} &= \gamma\frac{\sum_{i=1}^{n}\sqrt{H_{ii}\delta_i^2}}{n\left(1-p\right)}\sqrt{n}, \\ \lambda &= n\gamma^2\left(\frac{\sum_{i=1}^{n}\sqrt{H_{ii}\delta_i^2}}{n(1-p)}\right)^2 \end{aligned} \tag{A.16}$$

Substituting λ back, we have:

$$p_i = 1-\frac{n\left(1-p\right)\sqrt{H_{ii}\delta_i^2}}{\sum_{j=1}^{n}\sqrt{H_{jj}\delta_j^2}}. \tag{A.17}$$

and enforcing non-negativity finally yields the optimal solution:

$$p_i^* = max\left(0, 1-\frac{n(1-p)\sqrt{H_{ii}\delta_i^2}}{\sum_{j=1}^{n}\sqrt{H_{jj}\delta_j^2}}\right). \tag{A.18}$$

A.2. Proof of Theorem 2

Theorem 2 Given the constraint $\mathbb{E}\left[\left\|\Theta_{\text{fusion}}^{\ell}-\Theta_{\text{pre}}^{\ell}\right\|_1\right] < \eta$, there exists an upper bound for γ:

$$\gamma \leq \frac{\eta}{\sum_{i=1}^{n}|\delta_i|}, \tag{A.19}$$

where δ_i represents the i-th element of the delta parameters $\Delta\Theta^{\mathcal{T},\ell}$.

Proof. Let us expand the expected L1 norm:

$$\begin{aligned} &\mathbb{E}\left[\left\|\Theta_{\text{fusion}}^{\ell}-\Theta_{\text{pre}}^{\ell}\right\|_1\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n}\left|\Theta_{\text{fusion},i}^{\ell}-\Theta_{\text{pre},i}^{\ell}\right|\right] \\ &= \sum_{i=1}^{n}\mathbb{E}\left[\left|\Theta_{\text{fusion},i}^{\ell}-\Theta_{\text{pre},i}^{\ell}\right|\right] \\ &= \sum_{i=1}^{n}\left|p_i\Theta_{\text{pre},i}^{\ell}+\left(1-p_i\right)\left(\Theta_{\text{pre},i}^{\ell}+\frac{\gamma}{1-p_i}\delta_i\right)\right. \\ &\quad \left.-\Theta_{\text{pre},i}^{\ell}\right| \\ &= \sum_{i=1}^{n}\left|\gamma\delta_i\right| \end{aligned} \tag{A.20}$$

Given the constraint, we have:

$$\gamma \leq \frac{\eta}{\sum_{i=1}^{n}|\delta_i|} \tag{A.21}$$