Proposition 0.1. *Global Truncation-Aware truncation significantly enhances overall model accuracy.*

Proof. The global whitening matrix $S_k = X'_k X'^T_k$ exhibits controlled spectral properties, the effective condition number $\kappa(S_k)$ of the global whitening matrix is bounded by:

$$\kappa(S_k) \triangleq \frac{\sigma_{\max}(X'_k X'^T_k)}{\sigma_{\min}(X'_k X'^T_k)} \leq \frac{1 + \sum_{i=0}^{k-1} \alpha_i \|X_i X_i^T\|_2}{\sigma_{\min}(X_k X_k^T)} \triangleq \kappa_0 \tag{1}$$

where the decay coefficients α_i follow:

$$\alpha_i = 0.02 \log(1 + e^{(i+1)/N}) \tag{2}$$

The decay coefficients α_i ensures logarithmic decay of historical contributions, which implies:

$$\kappa(S_k)^{\text{global}} \ll \kappa(S_k)^{\text{local}} \quad \text{(Stabler gradients)} \tag{3}$$

Additionally, global features strictly dominate local features in mutual information, according to the condition number bound and cramer-rao inequality, we can get:

$$\frac{I(Y; W'_k X'_k)}{I(Y; W_k X_k)} \geq 1 + \underbrace{\frac{\sigma^2_{\min}(W'_k X'_k)}{\sigma^2_{\max}(W_k X_k)}}_{\text{Feature quality}} \cdot \underbrace{\frac{\kappa_0^{-1}}{1 - \kappa_0^{-1}}}_{\text{Stability benefit}} \tag{4}$$

Combining Feature Stability Condition and Information Preservation, we can get:

$$H(Y|W'_k X'_k) \leq H(Y|W_k X_k) - \log\left(1 + \frac{\sigma^2_{\min}(W'_k X'_k)}{\sigma^2_{\max}(W_k X_k)} \cdot \frac{1}{\kappa_0 - 1}\right) \tag{5}$$

The prediction error $P_e = 1 - A(M)$ then satisfies:

$$P_e^{\text{global}} \leq P_e^{\text{local}} - \underbrace{\frac{\log\left(1 + \frac{\sigma^2_{\min}}{\sigma^2_{\max}(\kappa_0 - 1)}\right)}{\log(|\mathcal{Y}| - 1)}}_{\Delta\text{(Explicit Improvement Term)}} \tag{6}$$

we derive the global accuracy superiority:

$$A(M)_{\text{global}} \geq A(M)_{\text{local}} + \underbrace{\log\left(1 + \frac{\sigma^2_{\min}}{\sigma^2_{\max}(\kappa_0 - 1)}\right)}_{\text{Explicit improvement term}} \frac{1}{\log(|\mathcal{Y}| - 1)} \tag{7}$$

This complete derivation demonstrates the direct role of global truncation-aware truncation in enhancing model accuracy through improved information preservation.

□

Proposition 0.2. *Momentum-Enhanced Alternating Least Squares achieve a lower loss*

Proof. We derive the momentum-enhanced alternating least squares method for optimizing the low-rank decomposition problem in MAS-SVD. The objective is to minimize the Frobenius norm between the original weight matrix W and its compressed counterpart W':

$$\mathcal{L}(U, V) = \|WX - UV^T\|_F^2 \tag{8}$$

where WX is the observed matrix, U and V are the low-rank factor matrices to be optimized. When only update U, the objective function is [**?**]:

$$\mathcal{L}(U_{t+1}, V_t) = \mathcal{L}(U_t, V_t) - \Delta_U \mathcal{L} \tag{9}$$

In contrary, we alternatively fix one variable and optimize the other, and the objective function is:

$$\mathcal{L}(U_{t+1}, V_{t+1}) = \mathcal{L}(U_t, V_t) - (\Delta_U \mathcal{L} + \Delta_V \mathcal{L}) \tag{10}$$

Since $\Delta_V \mathcal{L} > 0$, we have:

$$\mathcal{L}(U_{t+1}, V_{t+1}) < \mathcal{L}(U_{t+1}, V_t) \tag{11}$$

Similarly, when fix matrix U while update V, we have:

$$\mathcal{L}(U_{t+1}, V_{t+1}) < \mathcal{L}(U_t, V_{t+1}) \tag{12}$$

This shows alternating updates of U and V lead to a greater reduction in objective function. Additionally, we introduce momentum terms for U and V to enhance the performance of Alternating Least Squares. Among them, the momentum for U is updated as:

$$m_u = \beta m_u + (1 - \beta)\Delta U \tag{13}$$

where m_u is the momentum term for U, initialized as a zero matrix, β is the momentum coefficient (typically $\beta = 0.9$ or 0.95), and ΔU is the update for matrix U_{new} computed using the Alternating Least Squares update:

$$\Delta U = WXV(V^T V)^{-1} - U_{\text{old}} \tag{14}$$

where U_{old} is the matrix before update, The updated matrix U is then computed as:

$$U_{\text{new}} = U_{\text{old}} + \eta m_u \tag{15}$$

where η is the learning rate. Similarly, the momentum term for matrix V is updated as:

$$m_v = \beta m_v + (1 - \beta)\Delta V \tag{16}$$

where m_v is the momentum term for V, initialized as a zero matrix, and ΔV is the update for matrix V computed using the Alternating Least Squares update:

$$\Delta V = (U^T U)^{-1} U^T W X - V_{\text{old}} \tag{17}$$

Similarly, V_{old} is the matrix before update, The updated of matrix V_{new} is then computed as:

$$V_{\text{new}} = V_{\text{old}} + \eta m_v \tag{18}$$

The objective function at iteration t is $\mathcal{L}_t$. After applying the momentum-enhanced updates, the change in the objective function can be approximated using a Taylor expansion:

$$\mathcal{L}_{t+1} \approx \mathcal{L}_t - \eta \langle m_t, \nabla \mathcal{L}_t \rangle + O(\eta^2) \tag{19}$$

where $\langle \cdot, \cdot \rangle$ denotes the matrix inner product, m_t is the momentum term (either m_u or m_v), and $\nabla \mathcal{L}_t$ is the change of the objective function at iteration t. Under the Kurdyka-Łojasiewicz (KL) inequality, the momentum-enhanced ALS optimization satisfies:

$$\|\nabla \mathcal{L}(U_t, V_t)\|_2 \leq \frac{C}{(1-\beta)t} \tag{20}$$

where C depends on initial conditions and learning rate η, $\beta \in (0,1)$ is the momentum coefficient. While the vanilla ALS optimization satisfies:

$$\|\nabla \mathcal{L}(U_t, V_t)\|_2 \leq \frac{C}{t} \tag{21}$$

Hence, momentum-enhanced ALS achieves accelerated convergence compared to standard ALS. Additionally, The momentum update rule ensures the optimization avoids stationary points where:

$$\|\mathcal{L}\| \geq \frac{\beta}{1-\beta} \|m_{t-1}\| \tag{22}$$

This effectively filters out suboptimal critical points that would trap vanilla ALS. That is to say,When the update directions are consistent, the momentum term accumulates the update information, increasing the step size and accelerating the decrease of the loss. When the update directions are inconsistent, the momentum term smooths out the update direction, reducing oscillations in the loss function. Therefore, Momentum-enhanced Alternating Least Squares facilitates it easier to achieve a lower loss, avoiding over-adapting to local patterns and ensuring an optimized accuracy. □