

Appendix: Mixtures of Subspaces for Bandwidth Efficient Context Parallel Training

A Theoretical analysis

To provide formal support for our proposed joint optimization of the factorization $W = BUU^\top$ discussed in the main paper, we now present a rigorous analysis. Specifically, we consider the optimization dynamics on the product manifold $\mathcal{M} = \mathbb{R}^{d \times r} \times \text{St}(d, r)$, where B is optimized in standard Euclidean space and U resides on the Stiefel manifold. Under standard smoothness and boundedness assumptions commonly adopted in optimization theory, the following theorem establishes that our gradient descent updates converge linearly to a first-order stationary point.

Lemma 1. *Let $F : \mathbb{R}^{n \times n} \times \mathbb{R}^p \rightarrow \mathbb{R}$ be L -smooth and bounded below on rank- r matrices. Write each attention weight as $W = BUU^\top$ with $B \in \mathbb{R}^{n \times n}$ and $U \in \text{St}(n, r) := \{X \in \mathbb{R}^{n \times r} \mid X^\top X = I_r\}$, and collect the remaining parameters in $\vartheta \in \mathbb{R}^p$. Define $\Phi(B, U, \vartheta) := F(BUU^\top, \vartheta)$. Suppose Φ satisfies a PL inequality with constant $\mu > 0$ in a neighbourhood of an optimum $(B_\star, U_\star, \vartheta_\star)$ and that (B_0, U_0, ϑ_0) lies in this neighbourhood. Using the updates $B_{k+1} = B_k - \alpha_B \nabla_B \Phi_k$, $U_{k+1} = \exp_{U_k}(-\alpha_U \text{grad}_U \Phi_k)$, and $\vartheta_{k+1} = \vartheta_k - \alpha_\vartheta \nabla_\vartheta \Phi_k$ with fixed stepsizes $0 < \alpha_B, \alpha_U, \alpha_\vartheta \leq 1/L$ and $\alpha := \min\{\alpha_B, \alpha_U, \alpha_\vartheta\}$, we have for all $k \geq 0$*

$$\Phi(B_k, U_k, \vartheta_k) - \Phi_\star \leq (1 - \alpha\mu)^k [\Phi(B_0, U_0, \vartheta_0) - \Phi_\star], \quad \Phi_\star := \Phi(B_\star, U_\star, \vartheta_\star),$$

and the iterates converge linearly to $(B_\infty, U_\infty, \vartheta_\infty)$ with $B_\infty U_\infty U_\infty^\top = W_\star$ and $\vartheta_\infty = \vartheta_\star$.

Proof. Because F is L -smooth and $B \mapsto BUU^\top$ is linear,

$$\nabla_B \Phi(B, U, \vartheta) = [\nabla_W F(W, \vartheta)]_{W=BUU^\top} UU^\top.$$

And we also know $\|U\|$ is bounded so $\nabla_B \Phi(B, U, \vartheta)$ is L -Lipschitz in (B, U, ϑ) ; the same holds for $\nabla_\vartheta \Phi$. For $U \in \text{St}(n, r)$ the Riemannian gradient is

$$\text{grad}_U \Phi = \pi_{T_U \text{St}(n, r)}(\nabla_U \Phi),$$

Where T_U is the tangent space and π is the projection back on to $\text{St}(n, r)$. See that the projection $\pi_{T_U \text{St}(n, r)}$ is bounded, so $\text{grad}_U \Phi$ is also L -Lipschitz.

Now we consider gradient descent in each parameter block. By the Euclidean descent lemma [32],

$$\Phi(B_{k+1}, U_k, \vartheta_k) \leq \Phi_k - \frac{\alpha_B}{2} \|\nabla_B \Phi_k\|^2. \quad (1)$$

With B_{k+1}, ϑ_k fixed, the Riemannian descent lemma on $\text{St}(n, r)$ [2] gives

$$\Phi(B_{k+1}, U_{k+1}, \vartheta_k) \leq \Phi(B_{k+1}, U_k, \vartheta_k) - \frac{\alpha_U}{2} \|\nabla_U \Phi_k\|^2. \quad (2)$$

Finally, keeping (B_{k+1}, U_{k+1}) fixed,

$$\Phi_{k+1} \leq \Phi(B_{k+1}, U_{k+1}, \vartheta_k) - \frac{\alpha_\vartheta}{2} \|\nabla_\vartheta \Phi_k\|^2. \quad (3)$$

Summing (1)–(3) yields

$$\Phi_{k+1} \leq \Phi_k - \frac{1}{2} [\alpha_B \|\nabla_B \Phi_k\|^2 + \alpha_U \|\nabla_U \Phi_k\|^2 + \alpha_\vartheta \|\nabla_\vartheta \Phi_k\|^2]. \quad (4)$$

Let $\alpha := \min\{\alpha_B, \alpha_U, \alpha_\vartheta\}$.

By PL assumption, there exist $\mu > 0$ and a neighbourhood $\mathcal{N} \subseteq \mathbb{R}^{n \times n} \times \text{St}(n, r) \times \mathbb{R}^p$ containing the iterates such that

$$\|\nabla_B \Phi\|^2 + \|\nabla_U \Phi\|^2 + \|\nabla_\vartheta \Phi\|^2 \geq 2\mu [\Phi - \Phi_\star] \quad \forall (B, U, \vartheta) \in \mathcal{N}. \quad (5)$$

Combining (4) and (5) gives

$$\Phi_{k+1} - \Phi_\star \leq (1 - \alpha\mu) [\Phi_k - \Phi_\star],$$

and induction yields

$$\Phi_k - \Phi_* \leq (1 - \alpha\mu)^k [\Phi_0 - \Phi_*].$$

which is the claimed linear rate.

The geometric decay plus L -smoothness implies $\sum_k \|\nabla_B \Phi_k\|^2 < \infty$ (similarly for $\nabla_U \Phi_k$ and $\nabla_\theta \Phi_k$), hence all gradient blocks vanish. Any limit point $(B_\infty, U_\infty, \theta_\infty)$ therefore satisfies the first-order conditions and attains Φ_* . Consequently $W_k := B_k U_k U_k^\top$ converges to $W_* := B_\infty U_\infty U_\infty^\top$, and $\theta_k \rightarrow \theta_\infty = \theta_*$. \square

A.1 Geometric Impact of Rotational Reparameterisation

Next, to formally verify that our reparameterization $U(\theta) = R(\theta) \bar{U}$ preserves the stationary points of the optimization landscape, we present the complete theorem and proof in this appendix. Specifically, we rigorously demonstrate that optimizing in the unconstrained parameter space θ via ordinary SGD or Adam does not alter the nature or locations of local minima and strict saddle points of the original constrained optimization problem. This formalizes and extends the intuitive reasoning discussed in the main text, confirming that our rotational reparameterization is geometrically faithful and optimization-efficient.

Theorem 1. *Consider $\Phi(B, U, \vartheta)$ as defined in Proposition 1. and fix an orthonormal matrix $\bar{U} \in \text{St}(d, r)$. Suppose $\theta \mapsto R(\theta)$ is surjective (or dense) in $O(d) = \{R \in \mathbb{R}^{d \times d} | R^\top R = \mathbf{I}_d\}$. Define the Euclidean reparameterisation*

$$\hat{\Phi}(B, \theta, \vartheta) := \Phi(B, R(\theta) \bar{U} \bar{U}^\top R(\theta)^\top, \vartheta).$$

Then the sets of local minima of $\hat{\Phi}(B, \theta, \vartheta)$ and $\Phi(B, U, \vartheta)$ coincide.

Proof. Denote by $\phi(B, \theta, \vartheta) = (B, U, \vartheta)$ the *lifting map* from the re-parameterised domain $\mathcal{D}_1 = \mathbb{R}^{d \times d} \times \Theta$ to the original domain $\mathcal{D}_2 = \mathbb{R}^{d \times d} \times \text{St}(d, r)$, where $\theta \in \Theta$.

For any $U \in \text{St}(d, r)$ extend its columns to an orthogonal matrix $R \in O(d)$ with $UR^\top = \bar{U}$; surjectivity of $\theta \mapsto R(\theta)$ (or its denseness plus continuity) yields θ with $R(\theta) = R$, so $U = R(\theta) \bar{U}$. Hence $\phi(\mathcal{D}_1) = \mathcal{D}_2$. That is, ϕ is continuous and surjective.

Now, let $(B^*, \theta^*, \vartheta^*)$ be a local minimiser of $\hat{\Phi}$ and set $U^* := R(\theta^*) \bar{U} = \phi(B^*, \theta^*)_U$. Assume, towards a contradiction, that (B^*, U^*, ϑ^*) is *not* a local minimiser of Φ . Then there exists a sequence $(B_k, U_k) \rightarrow (B^*, U^*)$ with $\Phi(B_k, U_k, \vartheta_k) < \Phi(B^*, U^*, \vartheta^*)$. By surjectivity (or density) pick θ_k such that $U_k = R(\theta_k) \bar{U}$ and $\theta_k \rightarrow \theta^*$. Continuity of $\hat{\Phi} = \Phi \circ \phi$ gives $\hat{\Phi}(B_k, \theta_k, \vartheta_k) = \Phi(B_k, U_k, \vartheta_k) < \Phi(B^*, U^*, \vartheta^*) = \hat{\Phi}(B^*, \theta^*, \vartheta^*)$, contradicting local minimality of $\hat{\Phi}$. Therefore, local minima of $\hat{\Phi}$ lift to local minima of Φ . Let (B^*, θ^*)

Conversely, let $(\hat{B}, \hat{U}, \hat{\vartheta})$ be a local minimiser of Φ . Choose any $R \in O(d)$ with $\hat{U} = R \bar{U}$ and pick $\hat{\theta}$ such that $R(\hat{\theta}) = R$. If $(B, \theta, \vartheta) \rightarrow (\hat{B}, \hat{\theta}, \hat{\vartheta})$ then $\phi(B, \theta, \vartheta) \rightarrow (\hat{B}, \hat{U}, \hat{\vartheta})$, hence $\hat{\Phi}(B, \theta, \vartheta) = \Phi(B, U, \vartheta) \geq \Phi(\hat{B}, \hat{U}, \hat{\vartheta}) = \hat{\Phi}(\hat{B}, \hat{\theta}, \hat{\vartheta})$, so $(\hat{B}, \hat{\theta}, \hat{\vartheta})$ is a local minimiser of $\hat{\Phi}$.

Thus, ϕ induces a bijection between the sets of local minima of Φ and $\hat{\Phi}$; hence the two optimisation problems have exactly the same local minima. \square

Theorem 1 proves that the Euclidean reparameterisation $U(\theta) = R(\theta) \bar{U}$, $R(\theta) \in O(d)$ preserves *stationary points*. In this section we take a deeper look at how the transformation affects the *surrounding loss landscape*, focusing on curvature, conditioning, and global distortions. Specifically, we show that by constraining $\|\theta\|$ to a moderate range (via clipping), one can make sure that the curvature of the re-parameterized landscape remain similar to that of the original objective. Throughout, $\Phi(B, U, \vartheta)$ denotes the original objective, while $\hat{\Phi}(B, \theta, \vartheta) := \Phi(B, R(\theta) \bar{U}, \vartheta)$ is its pull-back to the unconstrained parameters θ .

A.1.1 Hessian pull-back formula

Let $g_U = \nabla_U \Phi \in \mathbb{R}^{d \times r}$, $H_U = \nabla_{UU}^2 \Phi \in \mathbb{R}^{dr \times dr}$ be the gradient and intrinsic (Euclidean) Hessian w.r.t. the Stiefel coordinates. Define the Jacobian of the reparameterisation $J(\theta) = \frac{\partial U(\theta)}{\partial \theta} \in \mathbb{R}^{dr \times m}$, $m = \dim \Theta$.

Lemma 2. For any θ the Euclidean Hessian of $\hat{\Phi}$ satisfies

$$H_\theta := \nabla_{\theta\theta}^2 \hat{\Phi} = J^\top H_U J + \sum_{i=1}^m (\partial_{\theta_i} J)^\top g_U. \quad (1)$$

At stationary points—i.e. when $g_U = 0$ —the second term vanishes and

$$H_\theta = J^\top H_U J. \quad (2)$$

Proof of Lemma 2. Fix (B, θ, ϑ) and denote³ $u(\theta) = \text{vec}(U(\theta))$, $g_U(u) = \nabla_u \Phi$, $J(\theta) = \frac{\partial u}{\partial \theta} \in \mathbb{R}^{dr \times m}$.

By the multivariate chain rule

$$\nabla_\theta \hat{\Phi} = J^\top g_U.$$

Differentiate once more:

$$H_\theta = \nabla_\theta (J^\top g_U) = (\nabla_\theta J^\top) g_U + J^\top (\nabla_\theta g_U).$$

Compute the two addends separately.

(i) *Derivative of the Jacobian.* Write the i -th column of J as $J_{(:,i)}$. Then $(\nabla_\theta J^\top) g_U = \sum_{i=1}^m (\partial_{\theta_i} J)^\top g_U$.

(ii) *Derivative of the pulled-back gradient.* Because g_U is a function of $u(\theta)$,

$$\nabla_\theta g_U = (\nabla_u g_U) \nabla_\theta u = H_U J.$$

Substituting (i) and (ii) yields

$$H_\theta = J^\top H_U J + \sum_{i=1}^m (\partial_{\theta_i} J)^\top g_U,$$

which is precisely (1). If θ is a stationary point of $\hat{\Phi}$, then $g_U = 0$ and the second term vanishes, giving (2). \square

Spectral consequences. Writing the singular values of J as $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ and the ordered eigenvalues of H_U as $\lambda_1 \geq \dots \geq \lambda_r$, classic congruence inequalities yield

$$\sigma_r^2 \lambda_r \leq \lambda_{\min}(H_\theta) \leq \lambda_{\max}(H_\theta) \leq \sigma_1^2 \lambda_1. \quad (3)$$

Hence the **condition number** transforms as $\kappa(H_\theta) \leq (\sigma_1/\sigma_r)^2 \kappa(H_U)$. Good scaling of J is therefore essential for maintaining a well-conditioned landscape.

A.1.2 Local properties with the exponential map

We instantiate $R(\theta)$ with the exponential map

$$R(\theta) = \exp\left(\sum_{i=1}^m \theta_i A_i\right), \quad A_i^\top = -A_i, \quad \langle A_i, A_j \rangle_F = \delta_{ij}, \quad (4)$$

³For clarity we vectorise matrices, writing $u = \text{vec}(U) \in \mathbb{R}^{dr}$, $g_U = \nabla_u \Phi \in \mathbb{R}^{dr}$, and $H_U = \nabla_{uu}^2 \Phi \in \mathbb{R}^{dr \times dr}$.

where $\{A_i\}$ is an *orthonormal basis* of the Lie algebra $\mathfrak{o}(d)$. For $\|\theta\| \ll 1$ the first-order expansion gives

$$R(\theta) = I + \sum_i \theta_i A_i + \mathcal{O}(\|\theta\|^2), \quad J_i := \frac{\partial U}{\partial \theta_i} = A_i \bar{U} + \mathcal{O}(\|\theta\|).$$

Because the $A_i \bar{U}$ are *orthonormal* in the embedded space $\mathbb{R}^{d \times r}$, the Gram matrix satisfies $J^\top J = I + \mathcal{O}(\|\theta\|)$. Hence near $\theta = 0$,

$$\sigma_1, \sigma_r = 1 + \mathcal{O}(\|\theta\|), \quad \kappa(J) = 1 + \mathcal{O}(\|\theta\|),$$

so by (3) the condition number of H_θ matches that of H_U up to first order. *Locally*, therefore, the exponential map is almost *isometric*: plain SGD or Adam on θ senses essentially the same curvature as a Riemannian optimiser on U .

Corollary 1. Fix any radius $\rho < \pi$ and let $\mathcal{B}_\rho = \{\theta \in \Theta \mid \|\theta\|_2 \leq \rho\}$. There exists $c_\rho > 0$ such that for all $\theta \in \mathcal{B}_\rho$ and stationary (B, θ, ϑ) ,

$$\frac{1}{1 + c_\rho} H_U \preceq H_\theta \preceq (1 + c_\rho) H_U.$$

In particular, $\kappa(H_\theta) \leq (1 + c_\rho)^2 \kappa(H_U)$.

Proof. Combine (2) with the bound $\|J^\top J - I\|_2 \leq c_\rho$ that follows from the expansion above and continuity of $R(\theta)$ inside \mathcal{B}_ρ . \square

Intuition. Corollary 1 asserts that, as long as the Lie-algebra parameter stays inside the ball $\mathcal{B}_\rho = \{\theta : \|\theta\|_2 \leq \rho\}$ with $\rho < \pi$, the Hessian in the new coordinates, H_θ , differs from the original Stiefel-space Hessian, H_U , by no more than a scalar factor $1 + c_\rho$ in either direction:

$$\frac{1}{1 + c_\rho} H_U \preceq H_\theta \preceq (1 + c_\rho) H_U, \quad c_\rho = \mathcal{O}(\rho).$$

Here “ \preceq ” denotes the usual Loewner order, so the inequality means that every quadratic form $v^\top H_\theta v$ lies between $v^\top H_U v / (1 + c_\rho)$ and $(1 + c_\rho) v^\top H_U v$. Consequently the condition number is inflated by at most the square of this factor, $\kappa(H_\theta) \leq (1 + c_\rho)^2 \kappa(H_U)$, and tends back to $\kappa(H_U)$ as $\rho \rightarrow 0$. In practical terms, when the rotation angles encoded by θ remain moderate ($\|\theta\| \lesssim 1$ rad), plain Euclidean optimisers experience almost the same curvature, and therefore the same step-size stability and convergence speed, as a Riemannian optimiser that works directly on U . Only as $\|\theta\|$ approaches the injectivity radius ($\approx \pi$) does the Jacobian cease to be nearly orthonormal, distorting the landscape enough to break this near-isometry. **In practice, we clip $\|\theta\| < 0.5$ to enforce this constraint.** Note that without this constraint, the models sometimes demonstrated unstable convergence.

A.1.3 Global distortions and injectivity radius

The exponential map is injective on $\|\theta\|_2 < \pi\sqrt{2}$ (the minimal distance to the cut-locus). As $\|\theta\| \rightarrow \pi$, several phenomena occur: **a) Jacobian degeneration.** Some singular values $\sigma_i(J)$ collapse to zero, flattening curvature along their directions and introducing plateaus in $\hat{\Phi}$ even if H_U is full-rank. **b) Anisotropic stretching.** Other singular values *blow up*, turning moderate curvature in H_U into steep walls in θ -space.

A.1.4 Practical takeaway

Rotational reparameterisation via the exponential map yields a *locally* isometric embedding of the Stiefel manifold into Euclidean space, ensuring that first-order methods experience essentially unchanged curvature near the solution set. However, as the Lie-algebra coordinates move towards the injectivity radius, the Jacobian may *distort* the landscape dramatically. Clipping or norm-projection of θ acts like a trust-region mechanism that retains favourable conditioning but risks biasing the search if the feasible clip radius is chosen too small.

The practitioner’s rule-of-thumb is therefore:

Maintain $\|\theta\|_2 \lesssim 1$ rad whenever possible; if larger rotations are essential, monitor $\|J\|_2$ or curvature statistics and re-centre / rescale when they inflate or collapse.

A.2 Rank collapse of the attention weights

To rigorously characterize the collapse of auxiliary projection heads onto data-dependent subspaces, as discussed in the main text, we present a detailed result and proof next. Specifically, we demonstrate formally how attention weights associated with infrequently activated directions shrink under optimization with weight decay. This confirms analytically that auxiliary projection heads become effectively redundant near convergence, justifying their safe removal and transition to a standard Transformer architecture without sacrificing accumulated predictive performance. First, we consider fixed projection matrices below. That is the case where U is not data dependent.

Proposition 1. *Let $U \in \mathbb{R}^{d \times r}$, $U^\top U = I_r$, $P := UU^\top$, $P_\perp := I_d - P$. Fix any ℓ_2 -regularised objective of the form $\mathcal{L}(W) = F(WP) + \frac{\lambda}{2} \|W\|_F^2$, $W \in \mathbb{R}^{d \times k}$, $\lambda > 0$. Assume only that $F : \mathbb{R}^{d \times k} \rightarrow \mathbb{R}$ is continuously differentiable. Consider the Gradient flow $\dot{W}(t) = -\nabla_W \mathcal{L}(W(t))$, Then the orthogonal component $W_\perp := WP_\perp$ obeys*

$$\|W_\perp(t)\|_F = (1 - \eta\lambda)^t \|W_\perp^{(0)}\|_F$$

Hence $W_\perp(t) \rightarrow 0$, and every limit point satisfies $W^* \in \text{col}(U)$.

Proof. Let $G := \nabla_X F(X)|_{X=WP} \in \mathbb{R}^{d \times k}$. Then we have,

$$\nabla_W F(WP) = GP. \quad (1)$$

Consider the full gradient of the regularized objective

$$\nabla_W \mathcal{L}(W) = GP + \lambda W. \quad (2)$$

Define the parallel and orthogonal blocks $W_\parallel := WP$, $W_\perp := WP_\perp$ so that $W = W_\parallel + W_\perp$. Because $PP_\perp = 0$,

$$W_\parallel P_\perp = 0, \quad W_\perp P = 0. \quad (3)$$

Using $\nabla_W \mathcal{L}(W) = GP + \lambda W$ in the update and projecting:

$$W^{(t+1)} P_\perp = (W^{(t)} - \eta(GP + \lambda W^{(t)})) P_\perp = (1 - \eta\lambda) W^{(t)} P_\perp.$$

Induction yields $\|W_\perp^{(t)}\|_F = (1 - \eta\lambda)^t \|W_\perp^{(0)}\|_F$.

So we have $\|W_\perp(t)\|_F \rightarrow 0$. Thus $P_\perp W(t) \rightarrow 0$, i.e. every accumulation point lies entirely in $\text{col}(U)$. \square

Next, we consider the case where U is data dependent under full-batch gradient descent.

Theorem 2. *For every sample x let $U(x) \in \mathbb{R}^{d \times r}$, $U(x)^\top U(x) = I_r$, $P(x) := U(x)U(x)^\top \in \mathbb{R}^{d \times d}$. Fix a loss family $F_x : \mathbb{R}^{d \times k} \rightarrow \mathbb{R}$ that is L bounded as: $\|\nabla_Z F_x(Z)\|_F \leq L \quad \forall x, Z$. Define the regularised objective $\mathcal{L}(W) := \mathbb{E}_x[F_x(WP(x))] + \frac{\lambda}{2} \|W\|_F^2$, $\lambda > 0$, $W \in \mathbb{R}^{d \times k}$. Let $Q \in \mathbb{R}^{d \times d}$ be an orthogonal projector and set $W_Q := WQ$. Introduce the average spectral overlap $p_Q := \mathbb{E}_x[\|P(x)Q\|_2] \in [0, 1]$. Run gradient descent $W^{(t+1)} = W^{(t)} - \eta \nabla_W \mathcal{L}(W^{(t)})$, $0 < \eta\lambda < 1$. Then for every $t \geq 0$ $\|W_Q^{(t)}\|_F \leq (1 - \eta\lambda)^t \|W_Q^{(0)}\|_F + \frac{p_Q L}{\lambda} [1 - (1 - \eta\lambda)^t]$ and consequently*

$$\limsup_{t \rightarrow \infty} \|W_Q^{(t)}\|_F \leq \frac{p_Q L}{\lambda}.$$

Proof. For $Z := WP(x)$ set $G(x) := \nabla_Z F_x(Z)$. Because $\nabla_W F_x(WP(x)) = G(x)P(x)$, adding the ℓ_2 -regulariser gives

$$\nabla_W \mathcal{L}(W) = \mathbb{E}_x[G(x)P(x)] + \lambda W.$$

Right-multiplying the GD update by Q ,

$$W_Q^{(t+1)} = (1 - \eta\lambda) W_Q^{(t)} - \eta \mathbb{E}_x[G(x)P(x)] Q.$$

Using the bound and $\|P(x)Q\|_2 \leq 1$,

$$\|G(x)P(x)Q\|_F \leq L \|P(x)Q\|_2.$$

Taking expectation and the definition of p_Q ,

$$\|\mathbb{E}_x[G(x)P(x)]Q\|_F \leq L p_Q.$$

Applying the bound in the recursion,

$$\|W_Q^{(t+1)}\|_F \leq (1 - \eta\lambda) \|W_Q^{(t)}\|_F + \eta L p_Q.$$

The inhomogeneous geometric series yields

$$\|W_Q^{(t)}\|_F \leq (1 - \eta\lambda)^t \|W_Q^{(0)}\|_F + \frac{L p_Q}{\lambda} [1 - (1 - \eta\lambda)^t].$$

Since $0 < 1 - \eta\lambda < 1$, $\lim_{t \rightarrow \infty} (1 - \eta\lambda)^t = 0$, giving $\limsup_{t \rightarrow \infty} \|W_Q^{(t)}\|_F \leq p_Q L / \lambda$. \square

Finally, we extend the above result to the case where U is data dependent and the networks is optimized via stochastic mini-batch gradient descent.

Theorem 3. For every sample x let $U(x) \in \mathbb{R}^{d \times r}$, $U(x)^\top U(x) = I_r$, $P(x) := U(x)U(x)^\top \in \mathbb{R}^{d \times d}$. Fix a loss family $F_x : \mathbb{R}^{d \times k} \rightarrow \mathbb{R}$ that is L bounded as: $\|\nabla_Z F_x(Z)\|_F \leq L \quad \forall x, Z$. Define the regularised objective $\mathcal{L}(W) := \mathbb{E}_x[F_x(WP(x))] + \frac{\lambda}{2} \|W\|_F^2$, $\lambda > 0$, $W \in \mathbb{R}^{d \times k}$. Let $Q \in \mathbb{R}^{d \times d}$ be an orthogonal projector and set $W_Q := WQ$. Introduce the average spectral overlap $p_Q := \mathbb{E}_x[\|P(x)Q\|_2] \in [0, 1]$. At iteration $t = 0, 1, \dots$ draw an i.i.d. sample x_t and perform

$$W^{(t+1)} = W^{(t)} - \eta_t g_t, \quad g_t := \nabla_W F_{x_t}(W^{(t)}P(x_t))P(x_t) + \lambda W^{(t)},$$

with stepsizes $\eta_t > 0$ satisfying $\eta_t \lambda < 1$. Set $W_Q^{(t)} := W^{(t)}Q$ and $p_Q := \mathbb{E}_x[\|P(x)Q\|_2] \in [0, 1]$.

If $\eta_t \equiv \eta$ and $0 < \eta\lambda < 1$, then for all $t \geq 0$

$$\mathbb{E}[\|W_Q^{(t)}\|_F] \leq (1 - \eta\lambda)^t \|W_Q^{(0)}\|_F + \frac{p_Q L}{\lambda} [1 - (1 - \eta\lambda)^t]$$

and hence $\limsup_{t \rightarrow \infty} \mathbb{E}[\|W_Q^{(t)}\|_F] \leq \frac{p_Q L}{\lambda}$.

Proof. Let $\mathcal{F}_t := \sigma(W^{(0)}, \dots, W^{(t)})$ be the natural filtration. All expectations $\mathbb{E}[\cdot]$ are taken over the drawn samples $\{x_s\}_{s \leq t}$.

Multiplying the update by Q on the right gives

$$W_Q^{(t+1)} = W_Q^{(t)} - \eta_t g_t Q.$$

Because g_t is an unbiased gradient estimate, $\mathbb{E}[g_t | \mathcal{F}_t] = \nabla_W \mathcal{L}(W^{(t)})$.

Take conditional expectation and use $\mathbb{E}[g_t Q | \mathcal{F}_t] = \mathbb{E}[\nabla_W F_{x_t}(\cdot)P(x_t)Q | \mathcal{F}_t] + \lambda W_Q^{(t)}$:

$$\mathbb{E}[W_Q^{(t+1)} | \mathcal{F}_t] = W_Q^{(t)} - \eta_t \lambda W_Q^{(t)} - \eta_t \mathbb{E}[\nabla_W F_{x_t}(\cdot)P(x_t)Q | \mathcal{F}_t].$$

Apply the Frobenius norm and the triangle inequality:

$$\mathbb{E}[\|W_Q^{(t+1)}\|_F \mid \mathcal{F}_t] \leq (1 - \eta_t \lambda) \|W_Q^{(t)}\|_F + \eta_t \left\| \mathbb{E}[\nabla_W F_{x_t}(\cdot) P(x_t) Q \mid \mathcal{F}_t] \right\|_F.$$

Since $\|\nabla_Z F_x(Z)\|_F \leq L$ for every x and Z ,

$$\|\nabla_W F_{x_t}(\cdot) P(x_t) Q\|_F \leq L \|P(x_t) Q\|_2.$$

Hence $\|\mathbb{E}[\nabla_W F_{x_t}(\cdot) P(x_t) Q \mid \mathcal{F}_t]\|_F \leq L p_Q$. Thus

$$\mathbb{E}[\|W_Q^{(t+1)}\|_F \mid \mathcal{F}_t] \leq (1 - \eta_t \lambda) \|W_Q^{(t)}\|_F + \eta_t p_Q L. \quad (5)$$

Let $\eta_t \equiv \eta$. Taking full expectation of (5) yields $y_{t+1} \leq (1 - \eta \lambda) y_t + \eta p_Q L$, where $y_t := \mathbb{E}[\|W_Q^{(t)}\|_F]$. Solving the linear recurrence gives

$$y_t \leq (1 - \eta \lambda)^t y_0 + \frac{\eta p_Q L}{\lambda} [1 - (1 - \eta \lambda)^t].$$

With $y_0 = \|W_Q^{(0)}\|_F$ we obtain the stated bound; letting $t \rightarrow \infty$ yields $\frac{p_Q L}{\lambda}$. □

B Attention output analysis

We investigate the attention output activation rank structure of pretrained frontier open-weight models, specifically focusing on prominent architectures such as LLaMA [47], Qwen [3], and Olmo [4] (Fig. 6, 7, and 8, respectively). Interestingly, our empirical analyses reveal a consistently observed low-rank structure across these diverse model families, suggesting that this phenomenon is intrinsic to transformer-based architectures rather than specific to certain model training procedures or datasets.

The low-rank behavior of attention outputs in transformers has garnered considerable interest, as it significantly impacts model efficiency, interpretability, and the potential for compression. Previous studies have documented similar findings; notably, Dong et al. [8] observed substantial rank reduction in the self-attention matrices of transformers during training, attributing it to implicit regularization effects induced by the training dynamics. Similarly, Abbe and colleagues [1] provided theoretical insights, demonstrating that rank collapse in attention mechanisms naturally arises from the iterative nature of gradient-based optimization processes.

Recent advancements in understanding transformer geometry and optimization further support our observations. Sanyal et al. [43] reported that transformers inherently favor lower-dimensional subspaces in their activations, leading to stable rank reduction, particularly in large-scale models. This intrinsic property has been leveraged in various model compression schemes, where exploiting the low-rank structure of attention outputs allows for significant reductions in memory footprint and computational overhead without adversely affecting model accuracy [14, 55].

Our findings confirm and extend these results by highlighting that the low-rank structure is not only prevalent across different model architectures but also robustly present across models trained on diverse datasets and with varying parameter scales. This observation underscores the potential universality of the low-rank phenomenon in transformer-based language models, further suggesting avenues for universally applicable model compression techniques and efficient parallelization strategies in decentralized training scenarios.

C Ablations

We conduct ablation studies across different network architectures to validate that the effectiveness of our compression scheme is not dependent on specific design choices.

Figures 9 and 10 show the convergence behavior of 8-layer and 32-layer models, respectively, both using 8 attention heads. Similarly, Figures 11 and 12 compare convergence between models with 8 and 24 attention heads, respectively; both models have 8 layers. All four configurations use an embedding dimension of 2048.

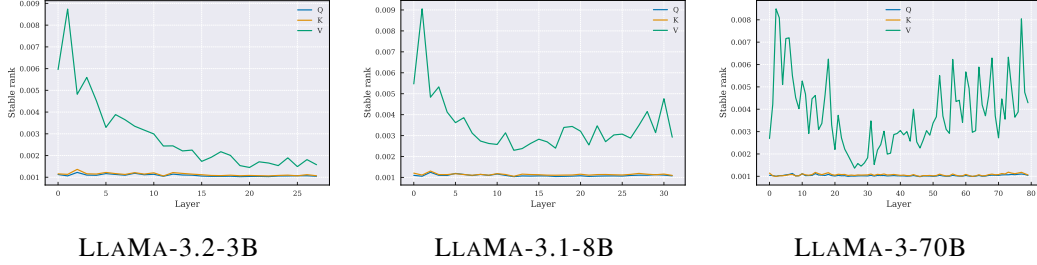


Figure 6: Stable rank distribution of the attention activations across LLAMA 3 models normalized by their maximum possible rank.

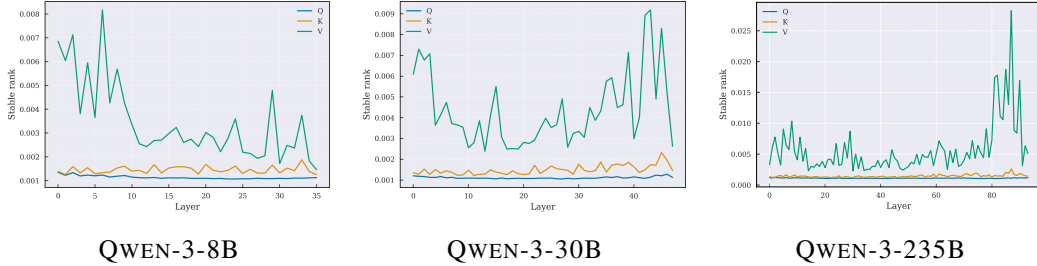


Figure 7: Stable rank distribution of the attention activations across QWEN 3 models normalized by their maximum possible rank.

Across all these variations, our compression method maintains convergence comparable to that of the centralized baseline, while achieving throughput competitive with centralized context-parallel models using 100Gbps connections—even though we only utilize 300Mbps connections.

In addition, we trained a model with an ultra-long context length of 256K tokens, distributed over 32 A100 GPUs. This model has 8 layers and 8 attention heads. As shown in Figure 14, even in this extreme setting, our compression technique enables decentralized training to match the convergence of the centralized baseline.

Note that we plot the loss curves against iterations (not wall-clock time) and hence, 300mbps decentralized (uncompressed) and 100Gbps centralized curves overlap. Throughput is reported next to the curves separately.

Table 5: Effect of Unplugging Step on Perplexity (PPL)

Unplugging Step	PPL
Not unplugged	22.64
5k	22.69
7k	22.64
8k	22.71
9k	22.71
9.5k	22.77
9.9k	23.12

C.1 Removing the projection components

In our experiments, we found that removing the projection components around the last 30%-5% range of training consistently results in stable performance across diverse settings. Based on this observation, we recommend using the last 30%-5% range of training as a simple and effective heuristic. Further, even when the projection components are removed earlier in training, the model is able to recover quickly, indicating a degree of robustness to the exact transition point. Nonetheless, the last 30% heuristic offers a practical, safer, and an automatable guideline. We present an ablation

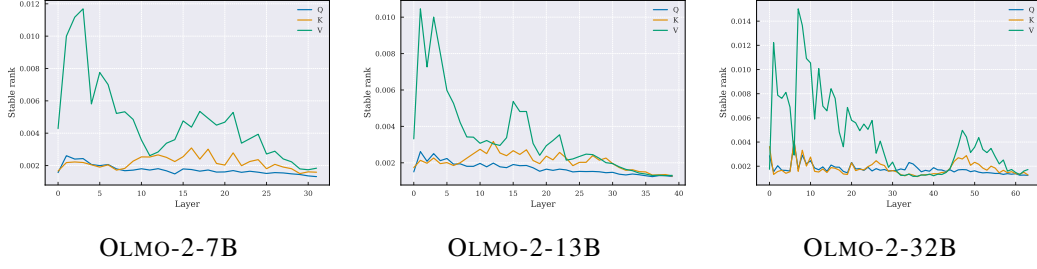


Figure 8: Stable rank distribution of the attention activations across OLMO 2 models normalized by their maximum possible rank.

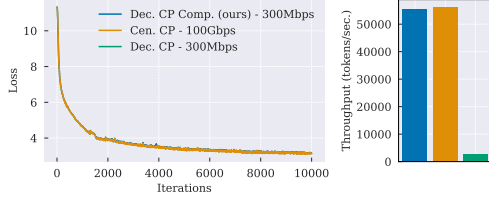


Figure 9: Loss over training iterations for 8-layer models (800M).

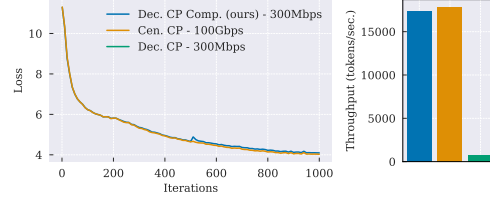


Figure 10: Loss over training iterations for 32-layer models (3B).

on FineWeb in Table 5 to further illustrate this (total steps 10k). Variations correspond to the third decimal point in the validation loss in most cases.

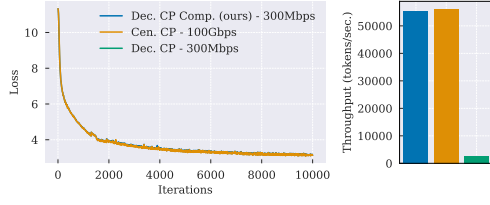


Figure 11: Loss over training iterations for 8-head models (800M).

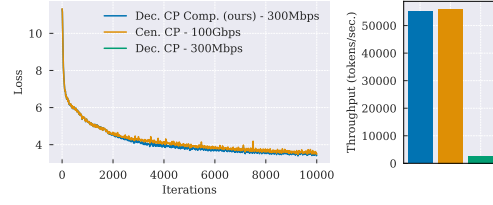


Figure 12: Loss over training iterations for 24-head models (2.5B).

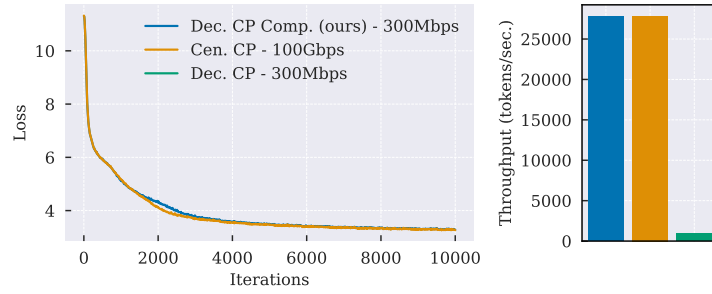


Figure 13: **Loss over training iterations for 256K context length training.** Our compression preserves the convergence with longer context lengths.

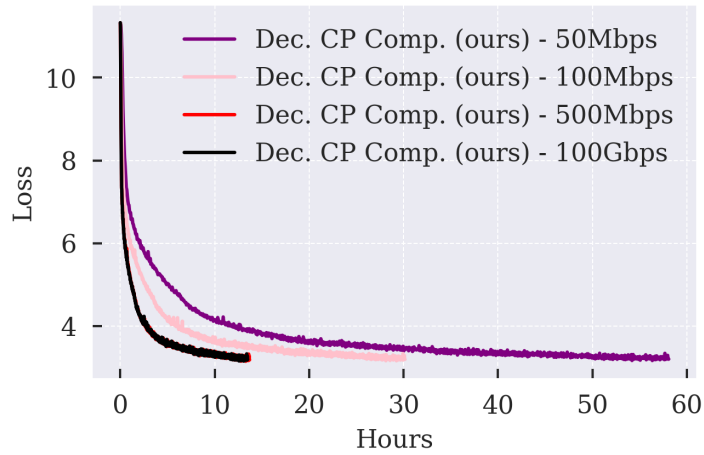


Figure 14: **Loss over wall-clock time across different bandwidths.** With our compression, the decentralized models achieve the convergence rate of centralized models at lower bandwidths.