Minimal Ranks, Maximum Confidence: Parameter-efficient Uncertainty Quantification for LoRA

Patryk Marszałek Klaudia Bałazy Jacek Tabor **Tomasz Kuśmierczyk**

tl;dr: Bayesian LoRA in a Subspace via SVD Projections

Method: B-LoRA-XS

- Compute **truncated SVD of** W^0: $W^0 \approx U_r S_r V_r^T$.
- Freeze projections $A = U_r S_r,\ B = V_r^T$.
- Insert small **adapter** R:

$$\Delta W = ARB, \quad R \in \mathbb{R}^{r \times r}.$$

- Learn **Bayesian posterior** over R

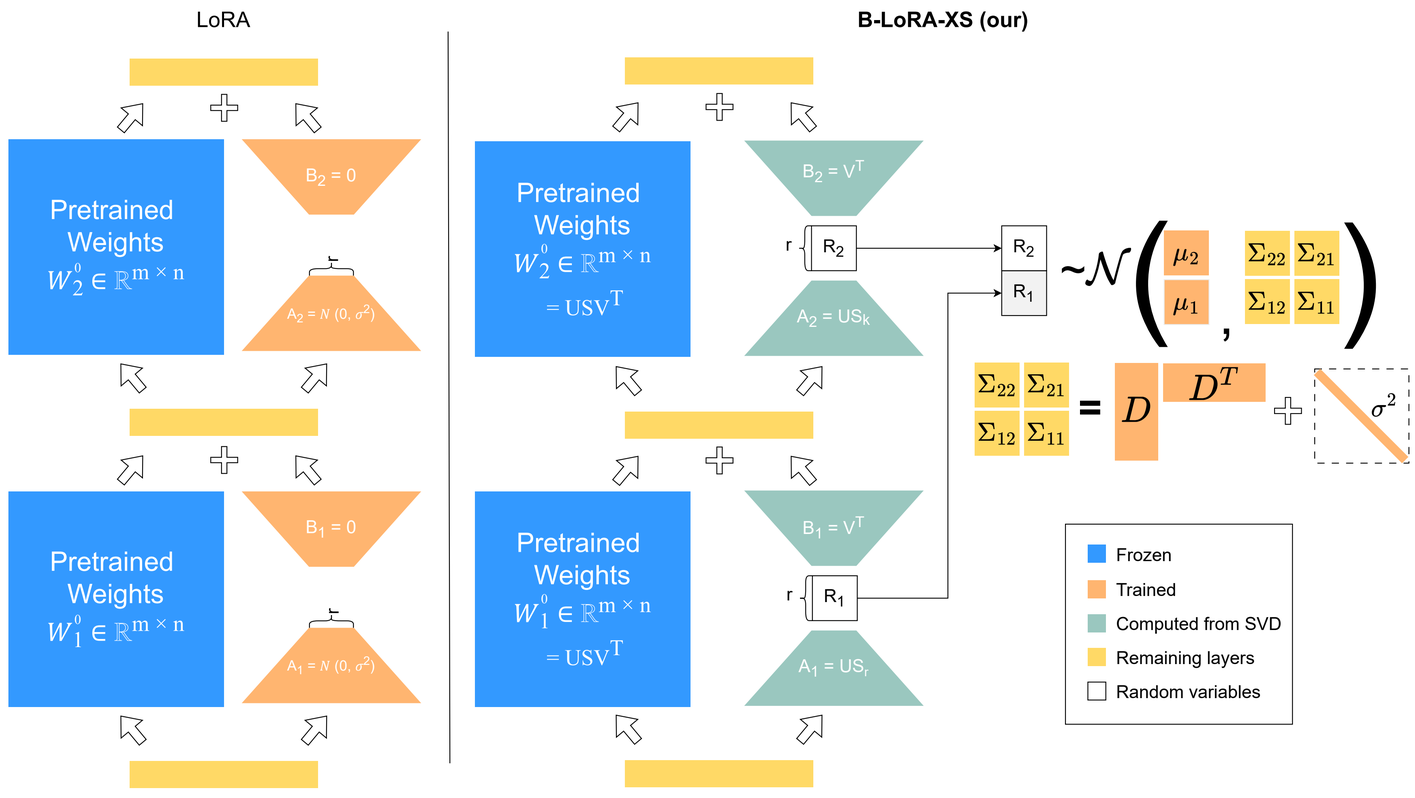
$$p(R \mid D) \approx \mathcal{N}(\mu_R, \Sigma_R),$$

with Σ_R low-rank (via SWAG).

Properties

- **Subspace learning:** inference restricted to the affine subspace $\mathcal{S}_B = \{w = w^0 + Pz | z \in \mathbb{R}^{r \times r}\}$, $w = \mathrm{vec}(W)$, $P = B^T \otimes A$.
- Covariance of ΔW induced by Kronecker structure:

$$\Sigma_{\Delta W} = (B^T \otimes A)\, \Sigma_R\, (B^T \otimes A)^\top.$$



Motivation

Achieve parameter efficiency while enabling epistemic uncertainty in LoRA

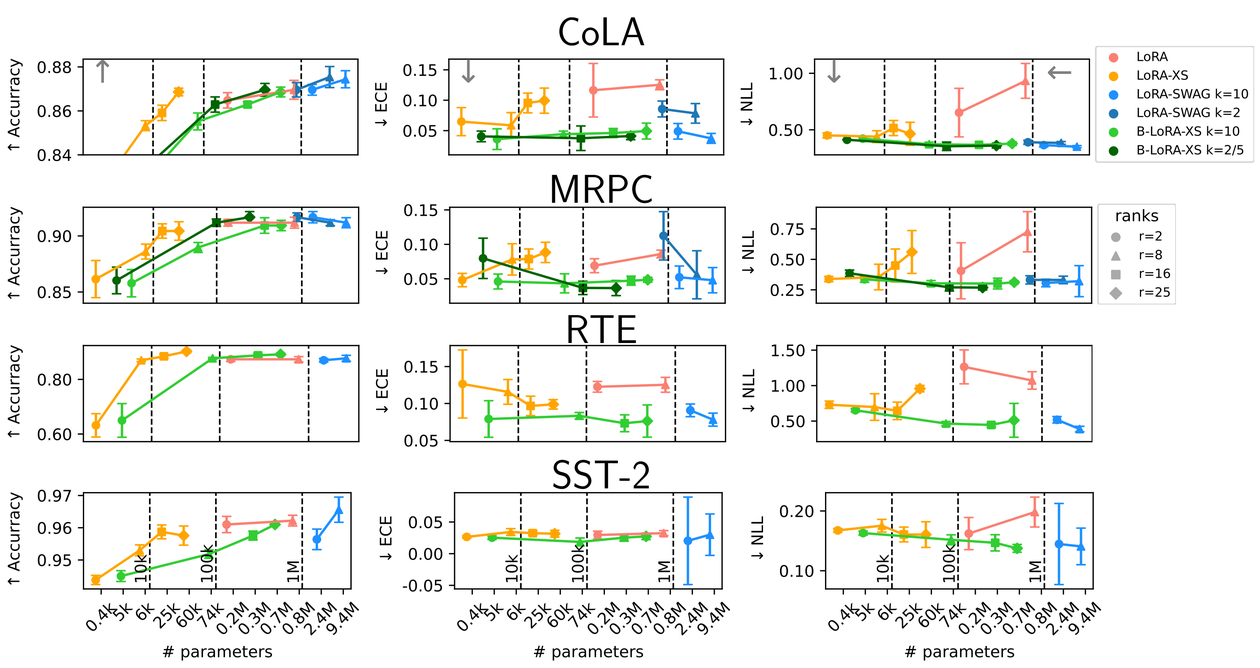
	LoRA	SVD-projected
Standard	No Uncertainty $n \times r \cdot (d_{\text{in}} + d_{\text{out}})$	No Uncertainty $n \times r^2$
Bayesian	Parameter Inefficient $n \times r \cdot (d_{\text{in}} + d_{\text{out}}) \cdot (k+2)$	B-LoRA-XS (Our) $(n \times r^2) \cdot (k+2)$

	Method	r	k	# Parameters
Standard	LoRA	2	-	0.2M
	LoRA	8	-	0.8M
	LoRA-XS	8	-	6k
	LoRA-XS	25	-	60k
Bayesian	SWAG-LoRA	2	10	2.4M
	SWAG-LoRA	8	10	9.4M
	SWAG-LoRA	8	5	5.5M
	Our	8	10	74k
	Our	25	10	0.7M
	Our	25	5	0.4M

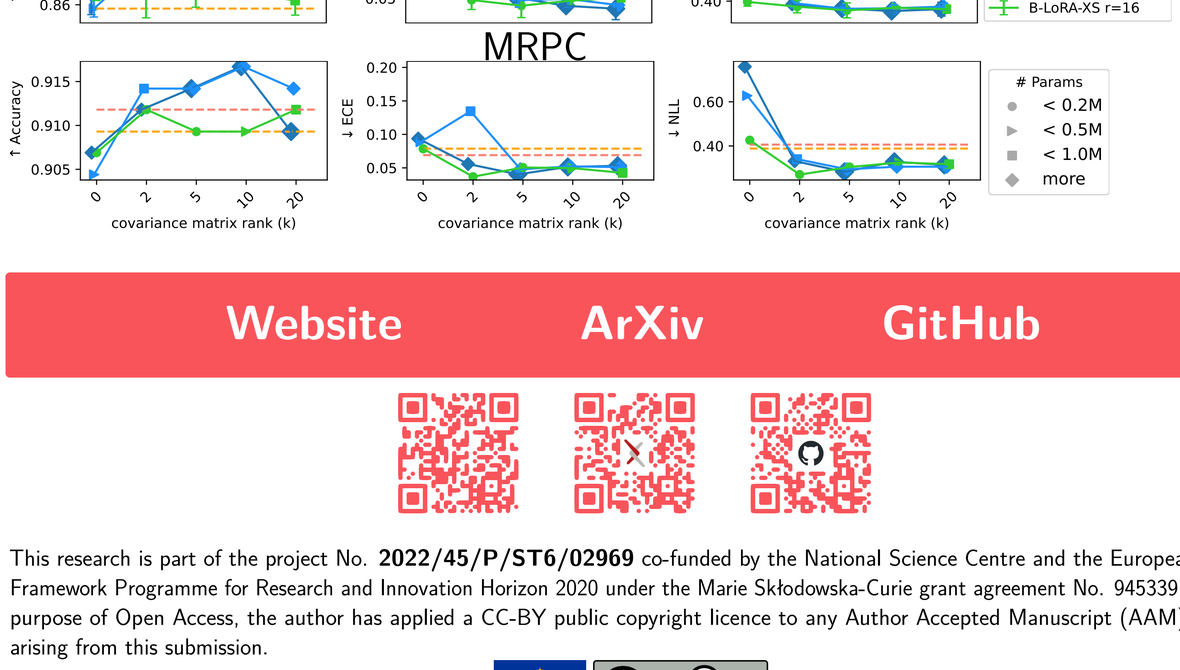
- r - LoRA rank
- k - covariance matrix rank
- n - number of LoRA layers

Performance at varying #Parameters

- *lower ECE and NLL* than standard LoRA across various parameter scales
- *fewer parameters* than Bayesian counterparts



Robustness against covariance rank k



Website **ArXiv** **GitHub**

This research is part of the project No. **2022/45/P/ST6/02969** co-funded by the National Science Centre and the European Union Framework Programme for Research and Innovation Horizon 2020 under the Marie Skłodowska-Curie grant agreement No. 945339.