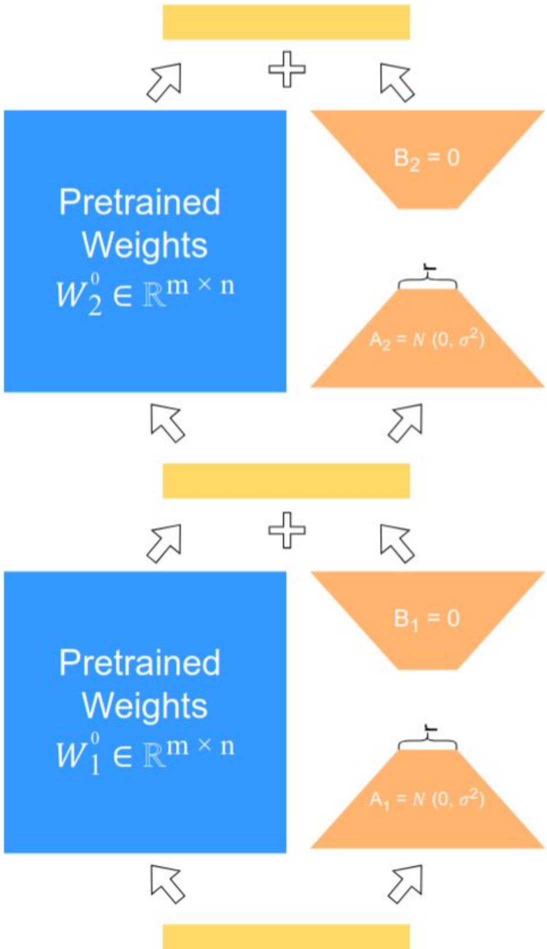
Minimal Ranks, Maximum Confidence: Parameter-efficient Uncertainty Quantification for LoRA

group of machine gmum learning research

JAGIELLONIAN UNIVERSITY IN KRAKÓW

Patryk Marszałek
Klaudia Bałazy
Jacek Tabor
Tomasz Kuśmierczyk

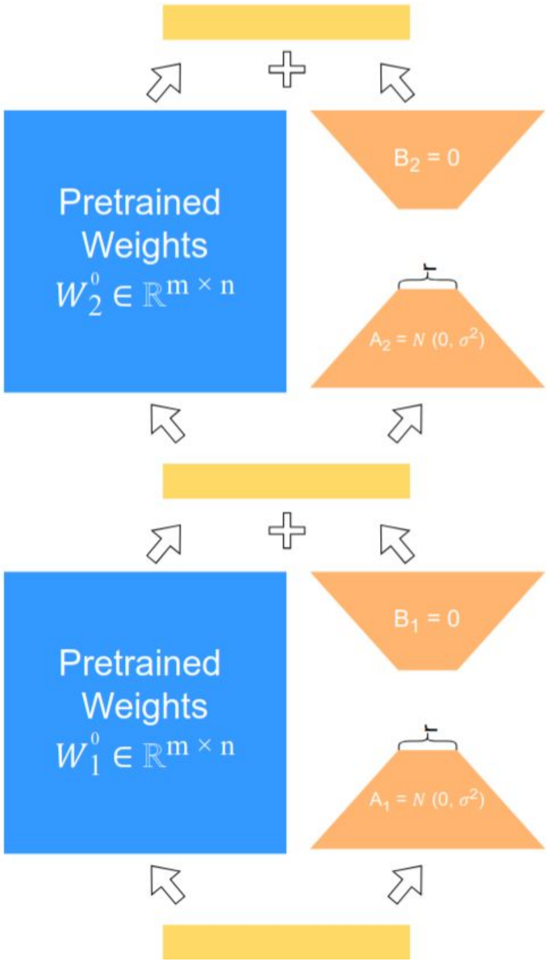
Standard LoRA



$$W = W^0 + \Delta W = W^0 + AB$$

	LoRA
Standard	No Uncertainty $n \times r \cdot (d_{\text{in}} + d_{\text{out}})$

Bayesian LoRA

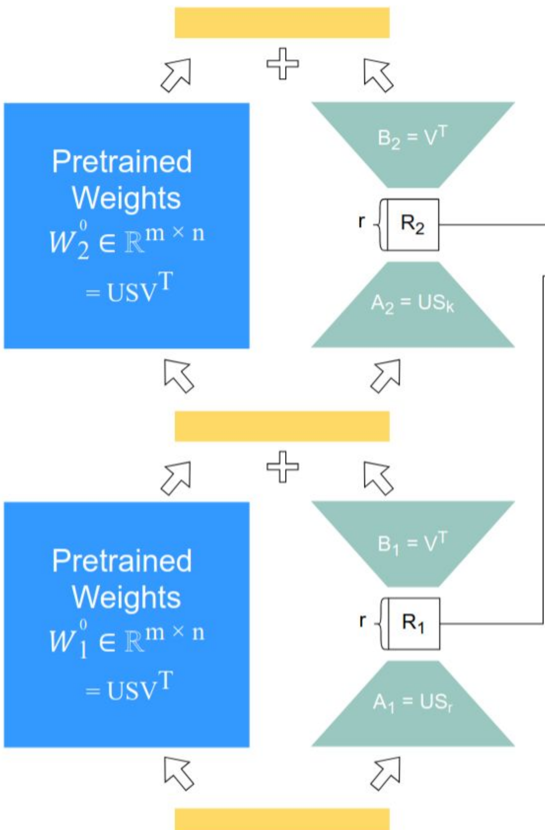


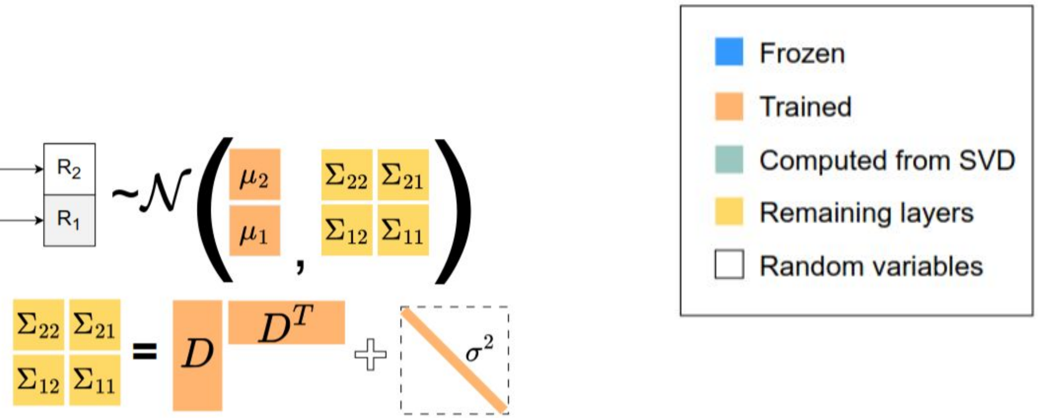
$$W = W^0 + \Delta W = W^0 + AB$$

$$[A, B] \sim \mathcal{N}(\mu, \Sigma)$$

	LoRA
Standard	No Uncertainty $n \times r \cdot (d_{\text{in}} + d_{\text{out}})$
Bayesian	Parameter Inefficient $n \times r \cdot (d_{\text{in}} + d_{\text{out}}) \cdot (k+2)$

B-LoRA-XS: epistemic uncertainty and parameter efficiency





	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) $(n \times r^2) \cdot (k+2)$

B-LoRA-XS: Algorithm

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

B-LoRA-XS: Algorithm

- 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:

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

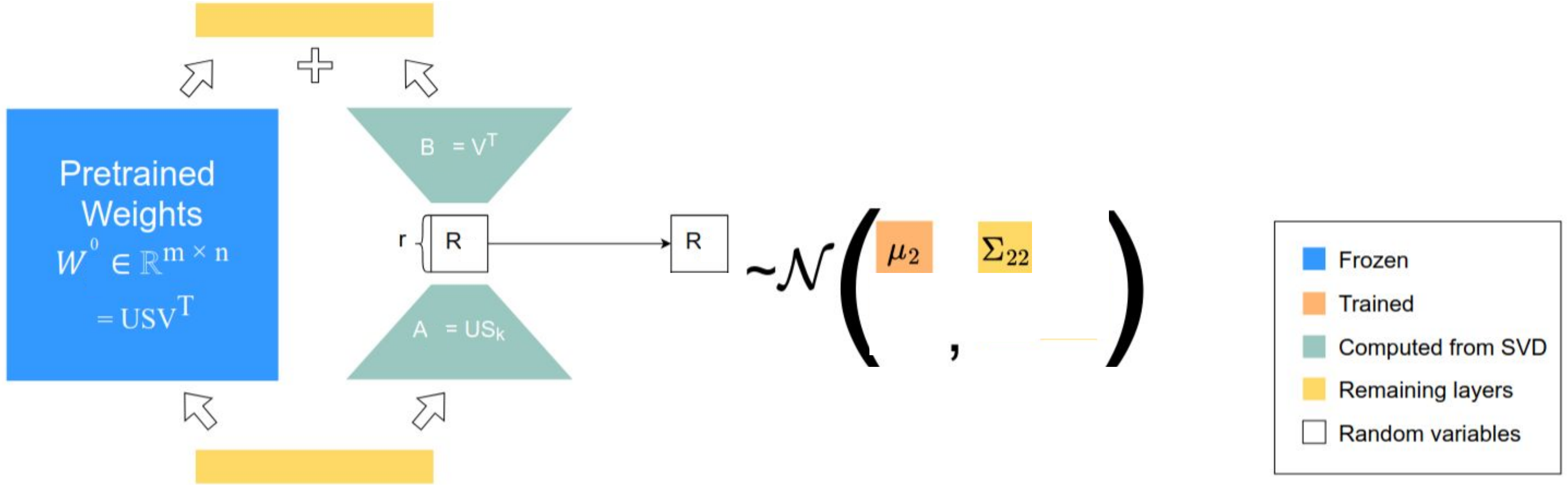
B-LoRA-XS: Algorithm

- 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

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



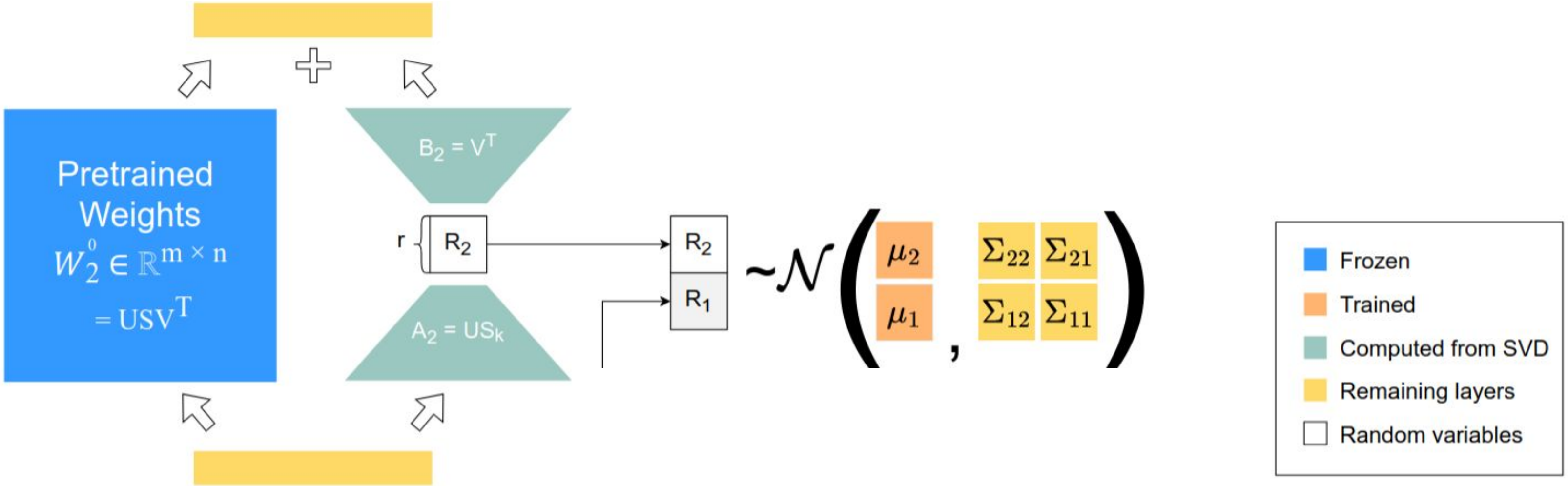
B-LoRA-XS: Algorithm

- 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

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



B-LoRA-XS: Algorithm

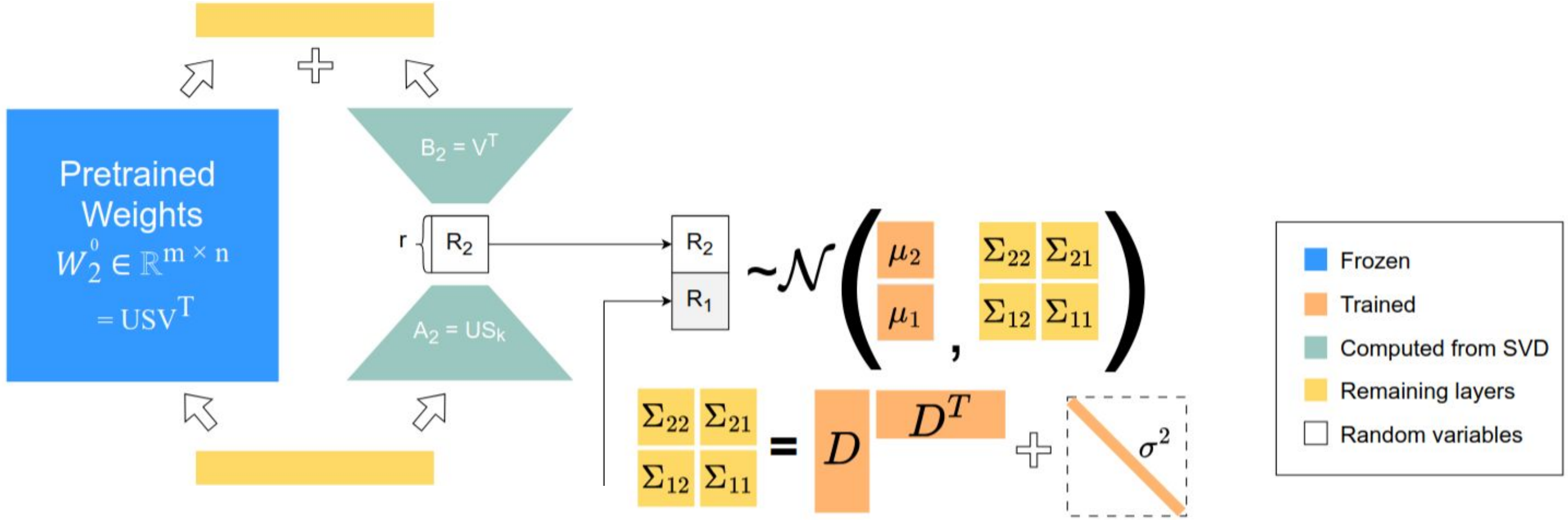
- Compute **truncated SVD of** W^0: $W^0 \approx U_r S_r V_r^T$.
- Freeze projections $A = U_r S_r,\ B = V_r^T$.
- Insert small **adapter** R:

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

- Learn **Bayesian posterior** over R

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

with Σ_R low-rank (via SWAG).

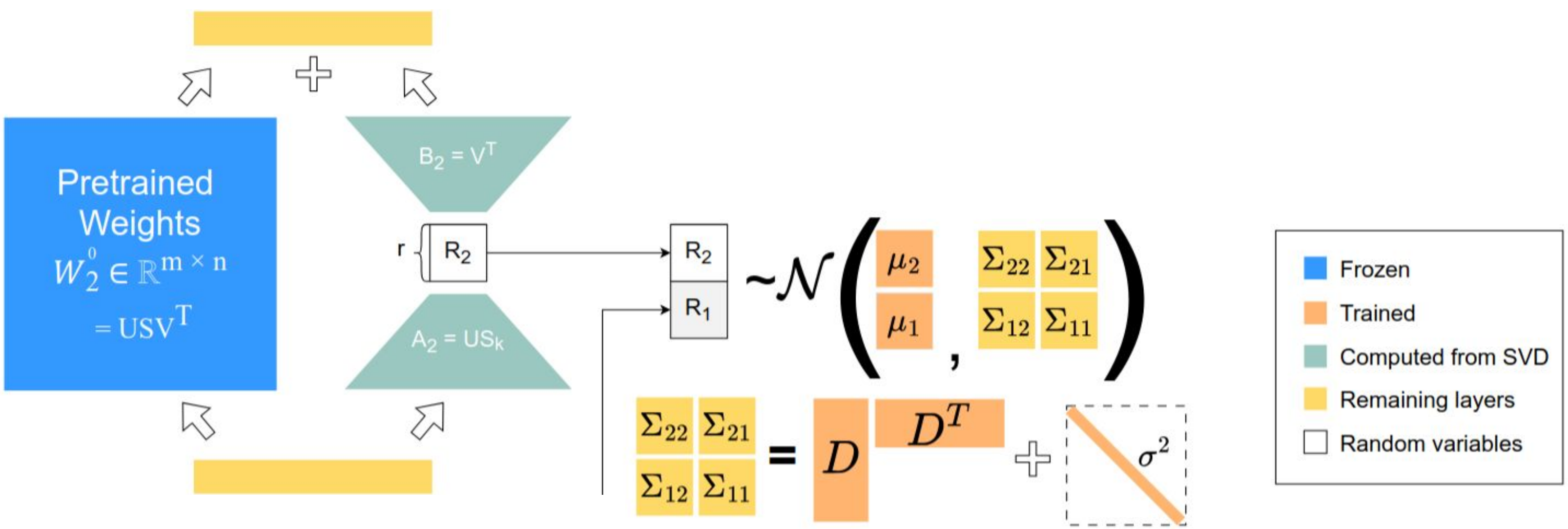


B-LoRA-XS: Properties

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

Subspace Inference for Bayesian Deep Learning

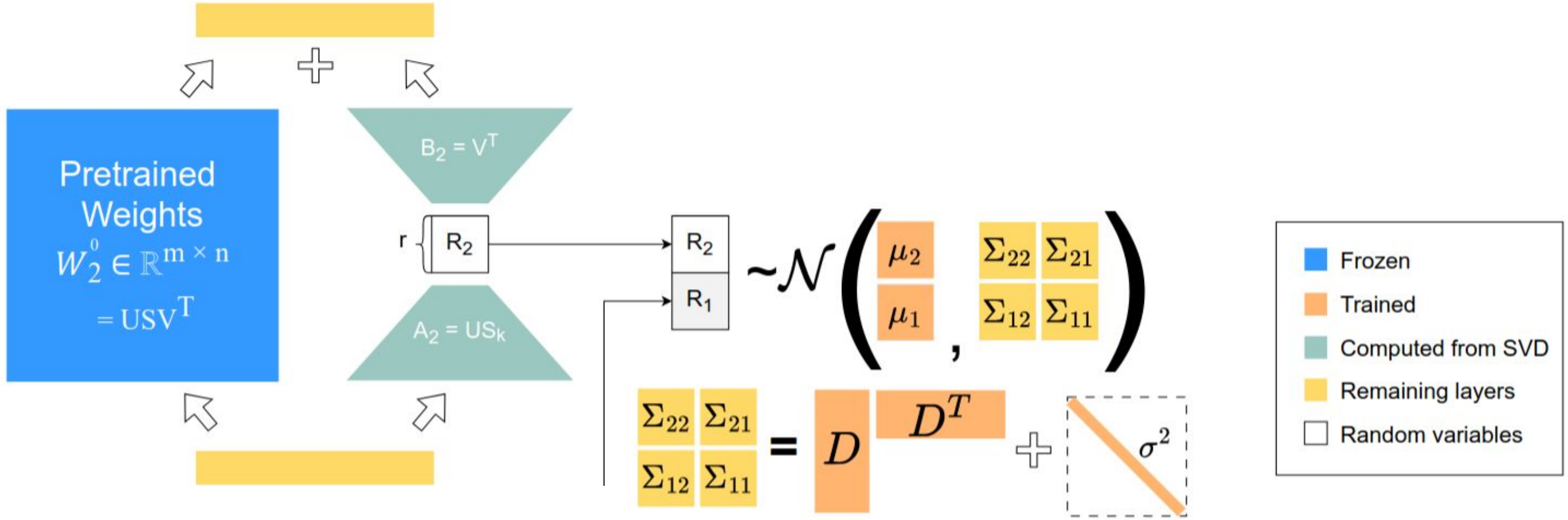
Pavel Izmailov, Wesley J. Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, Andrew Gordon Wilson *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, PMLR 115:1169-1179, 2020.



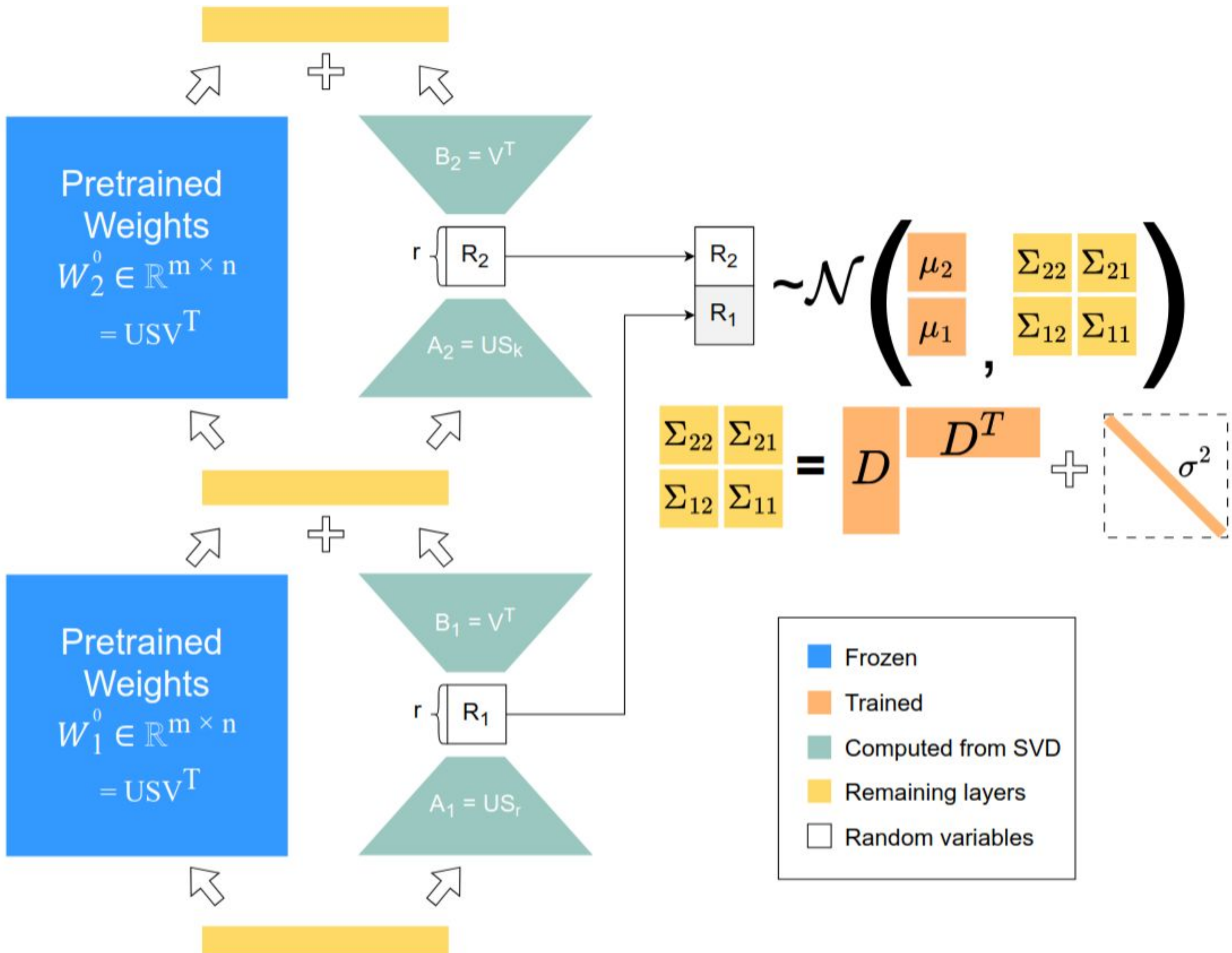
B-LoRA-XS: Properties

- **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:

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

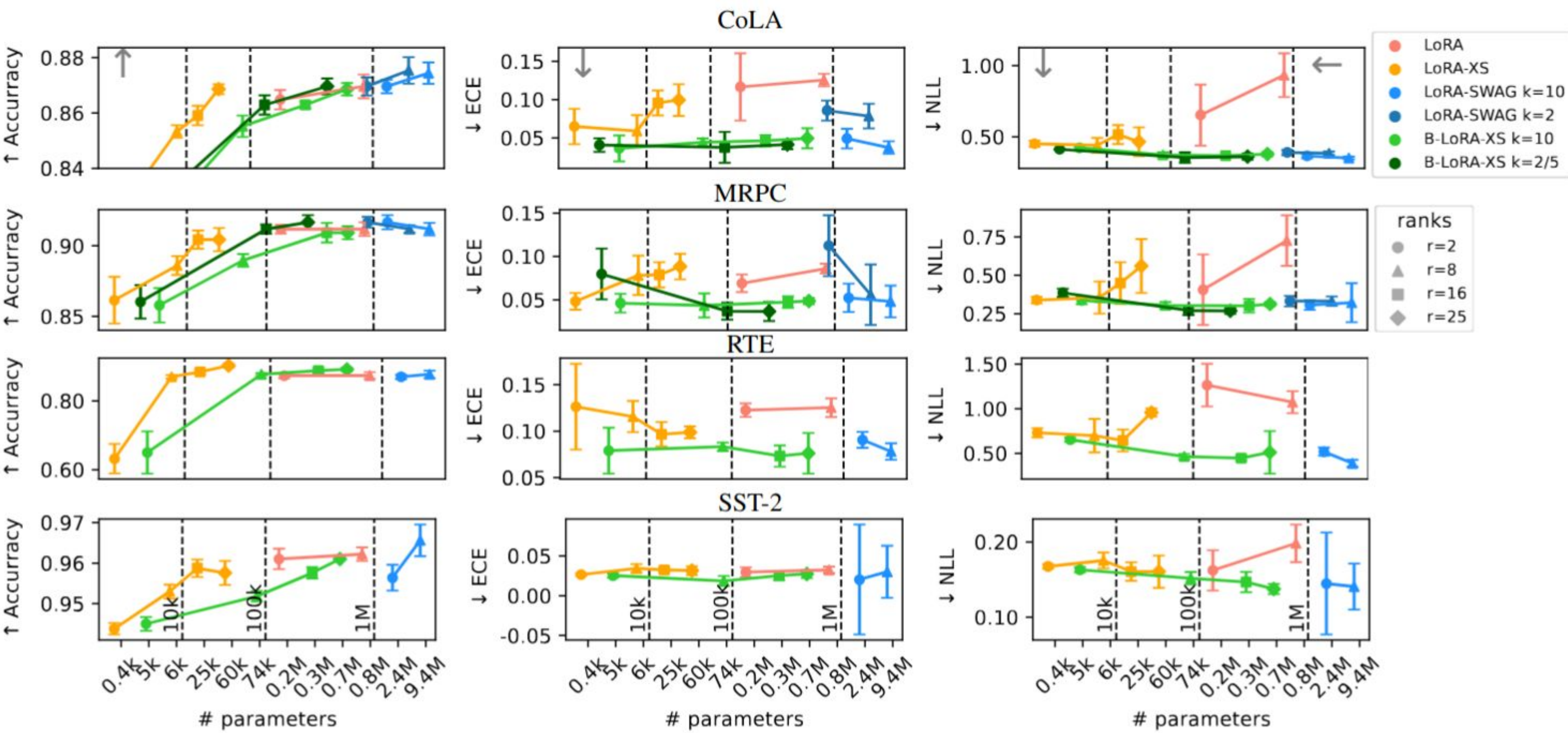


B-LoRA-XS

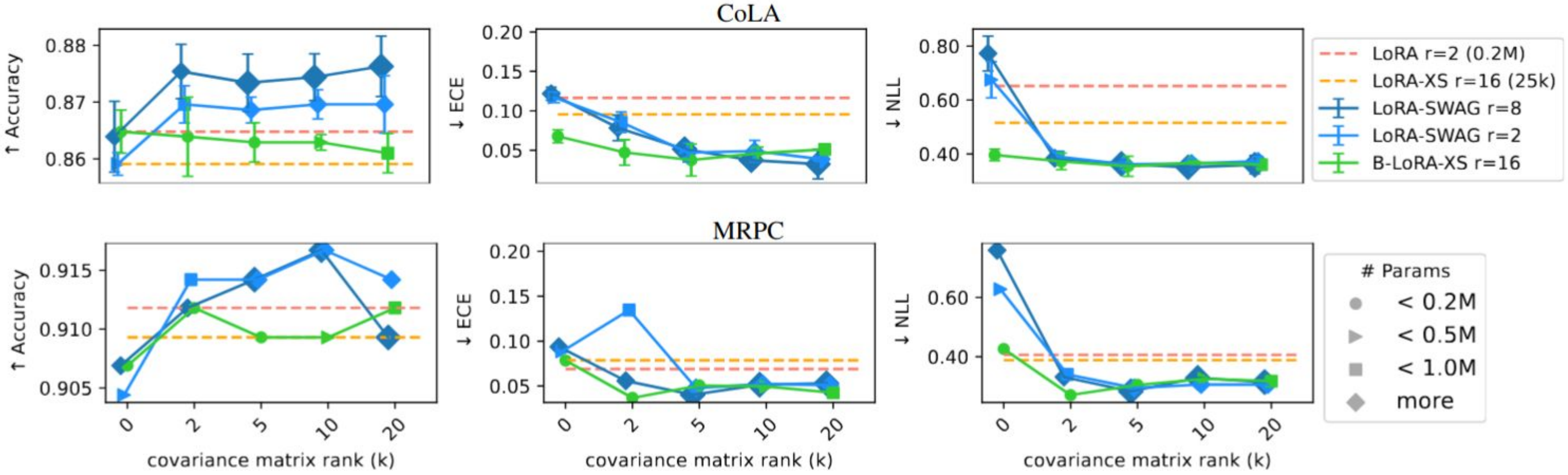


Pretrained Weights
$W_2^0 \in \mathbb{R}^{m \times n} = USV^T$
$B_2 = V^T$
r
R_2
$A_2 = US_k$
R_2
R_1
$\sim \mathcal{N}\left(\begin{matrix}\mu_2\\ \mu_1\end{matrix}, \begin{matrix}\Sigma_{22} & \Sigma_{21}\\ \Sigma_{12} & \Sigma_{11}\end{matrix}\right)$
$\begin{matrix}\Sigma_{22} & \Sigma_{21}\\ \Sigma_{12} & \Sigma_{11}\end{matrix} = D\,D^T + \sigma^2$
Pretrained Weights
$W_1^0 \in \mathbb{R}^{m \times n} = USV^T$
$B_1 = V^T$
r
R_1
$A_1 = US_r$
Frozen
Trained
Computed from SVD
Remaining layers
Random variables

Results: Accuracy / ECE / NLL vs #Params



Ablation: covariance rank k



https://bayes.ii.uj.edu.pl

https://arxiv.org/abs/2502.12122

*Badania zrealizowane w ramach projektu nr **2022/45/P/ST6/02969** współfinansowanego ze środków Narodowego Centrum Nauki oraz programu ramowego Unii Europejskiej w zakresie badań naukowych i innowacji Horyzont 2020 na podstawie umowy nr 945339 w ramach działań „Marie Skłodowska-Curie". Dla celów Open Access autor udostępnia na licencji CC-BY (Creative Commons – uznanie autorstwa) każdą wersję AAM, która może powstać w oparciu o niniejszy manuskrypt;*