

393 7 Algorithms

Algorithm 3: The functions `basis_augmentation`, and `truncation` of the used algorithms.

```

1 def basis_augmentation( $B$ : old basis,  $G_B$ : basis dynamics):
2    $\hat{B} \leftarrow \text{ortho}([G_B \mid B])$           /* orthonormalization, e.g. Gram-Schmidt */
3   return  $\hat{B}$ 
4 def truncation( $\hat{S}$ : augmented coefficient,  $\hat{S}_V$ : augmented momentum,  $\hat{U}$ : augmented basis,
    $\hat{V}$ : augmented co-basis ):
5    $P_{r_1}, \Sigma_{r_1}, Q_{r_1} \leftarrow \text{truncated svd}(\hat{S})$  with threshold  $\vartheta$  to new rank  $r_1$ 
6    $U \leftarrow \hat{U} P_{r_1}; V \leftarrow \hat{V} Q_{r_1}$                                 /* Basis update */
7    $S \leftarrow \Sigma_{r_1}; S_V \leftarrow U^\top \hat{U} \hat{S}_V \hat{V}^\top V$           /* Coefficient update */
8   return  $U, S, V, S_V$ 
9 def truncation( $\hat{S}$ : augmented coefficient,  $\hat{S}_V$ : augmented momentum,  $\hat{S}_K$ : augmented 2nd
   momentum,  $\hat{U}$ : augmented basis,  $\hat{V}$ : augmented co-basis ):
10   $P_{r_1}, \Sigma_{r_1}, Q_{r_1} \leftarrow \text{truncated svd}(\hat{S})$  with threshold  $\vartheta$  to new rank  $r_1$ 
11   $U \leftarrow \hat{U} P_{r_1}; V \leftarrow \hat{V} Q_{r_1}$                                 /* Basis update */
12   $S \leftarrow \Sigma_{r_1}; S_V \leftarrow U^\top \hat{U} \hat{S}_V \hat{V}^\top V; S_K \leftarrow U^\top \hat{U} \hat{S}_K \hat{V}^\top V$  /* Coefficient update */
13  return  $U, S, V, S_V, S_K$ 

```

Algorithm 4: Single iteration of the (full-rank version of) Adam.

Input : Initial parameter vector $W \in \mathbb{R}^{n \times n}$;
 \mathcal{V} : Initial 1st moment;
 \mathcal{K} : Initial 2nd moment;
Gradient $g = \nabla_W \mathcal{L}(W)$;
 λ : learning rate;
 β_1, β_2 : Adam momentum parameters;
 ϵ : Small stability constant.

```

1 Evaluate  $\mathcal{L}(W)$ 
2  $g \leftarrow \nabla_W \mathcal{L}(W)$                                 /* Compute gradient */
3  $\mathcal{V} \leftarrow \beta_1 \mathcal{V} + (1 - \beta_1)g$                   /* 1st moment estimate */
4  $\mathcal{K} \leftarrow \beta_2 \mathcal{K} + (1 - \beta_2)g^2$                 /* 2nd moment estimate (element-wise square) */
5  $\hat{\mathcal{V}} \leftarrow \frac{\mathcal{V}}{1 - \beta_1^t}, \quad \hat{\mathcal{K}} \leftarrow \frac{\mathcal{K}}{1 - \beta_2^t}$           /* Bias correction */
6  $W \leftarrow W - \lambda \frac{\hat{\mathcal{V}}}{\sqrt{\hat{\mathcal{K}} + \epsilon}}$                 /* Parameter update */

```

394 8 Numerical Analysis

395 **Theorem 1** (Convergence). Let $W(t)$ be the solution of (4) and let \mathcal{L} be bounded from below. Then,
396 $W(t)$ converges to a W^* which fulfills the low-rank optimality condition

$$P(W^*) \nabla_W \mathcal{L}(W^*) = 0. \quad (6)$$

397 *Proof.* Let us define the energy as

$$E(t) := \mathcal{L}(W(t)) + \frac{1}{2} \|\mathcal{V}(t)\|^2.$$

398 The time derivative is given by

$$\begin{aligned} \dot{E}(t) &:= \langle \nabla_W \mathcal{L}(W(t)), \dot{W}(t) \rangle + \langle \mathcal{V}(t), \dot{\mathcal{V}}(t) \rangle \\ &= \langle \nabla_W \mathcal{L}(W(t)), P(W(t)) \mathcal{V}(t) \rangle + \langle \mathcal{V}(t), -\gamma \mathcal{V}(t) - P(W(t)) \nabla_W \mathcal{L}(W(t)) \rangle. \end{aligned}$$

399 Since P is self-adjoint this directly gives

$$\begin{aligned}\dot{E}(t) &= \langle P(W(t)) \nabla_W \mathcal{L}(W(t)), \mathcal{V}(t) \rangle + \langle \mathcal{V}(t), -\gamma \mathcal{V}(t) - P(W(t)) \nabla_W \mathcal{L}(W(t)) \rangle \\ &= -\gamma \|\mathcal{V}(t)\|^2.\end{aligned}$$

400 Hence, if \mathcal{L} is bounded from below, this means that $\lim_{t \rightarrow \infty} E(t) = E_\infty$ with E_∞ finite and

$$E_\infty = E(0) - \gamma \int_0^\infty \|\mathcal{V}(t)\|^2 dt.$$

401 This implies that $\lim_{t \rightarrow \infty} \mathcal{V}(t) = 0$ and thus $\lim_{t \rightarrow \infty} \dot{W}(t) = \lim_{t \rightarrow \infty} P(W(t)) \mathcal{V}(t) = 0$. Hence,
402 since $\mathcal{V}(t), W(t)$ converge to a steady state and $\lim_{t \rightarrow \infty} \mathcal{V}(t) = 0$, the evolution equation for \mathcal{V} gives
403 $P(W(t)) \nabla_W \mathcal{L}(W(t)) = 0$ as $t \rightarrow \infty$. \square

404 We can obtain a similar, but not equivalent, result when solving a low-rank gradient flow of the form
405 5 instead:

406 **Theorem 2** (Convergence of low-rank factors). *The low-rank gradient flow*

$$\dot{U}_v = -(I - U_v U_v^\top) \nabla_W \mathcal{L} V_v S_v^{-1}, \quad (7a)$$

$$\dot{V}_v = -(I - V_v V_v^\top) \nabla_W \mathcal{L}^\top U_v S_v^{-\top}, \quad (7b)$$

$$\dot{S}_v = -\gamma S_v - U_v^\top \nabla_W \mathcal{L} V_v, \quad (7c)$$

407 fulfills

$$\dot{\mathcal{V}} = -\gamma \mathcal{V} - P(\mathcal{V}) \nabla_W \mathcal{L}.$$

408 *Proof.* By the product rule we have

$$\begin{aligned}\dot{\mathcal{V}} &= \dot{U}_v S_v V_v^\top + U_v \dot{S}_v V_v^\top + U_v S_v \dot{V}_v^\top \\ &= -\gamma \mathcal{V} - (I - U_v U_v^\top) \nabla_W \mathcal{L} V_v V_v^\top - U_v U_v^\top \nabla_W \mathcal{L} V_v V_v^\top - U_v U_v^\top \nabla_W \mathcal{L} (I - V_v V_v^\top) \\ &= -\gamma \mathcal{V} - P(\mathcal{V}) \nabla_W \mathcal{L}.\end{aligned} \quad (8)$$

409 \square

410 **Theorem 3** (Error-bound). *For an integer k , let $t = k\lambda$. Let $W(t)$ be the solution of 4, and let W_t^r ,
411 \mathcal{V}_t be the factorized low-rank solution after k steps with Algorithm 1. Assume that for any $Z \in \mathcal{M}_r$ in
412 a neighborhood of W_t^r , we have $\|(I - P(Z)) \nabla \mathcal{L}(Z)\| < \varepsilon$ and $\|\hat{U}_t \hat{U}_t^\top \mathcal{V}_t \hat{V}_t \hat{V}_t^\top - U_t U_t^\top \mathcal{V}_t V_t V_t^\top\| \leq$
413 $\hat{\vartheta}$, where $\|\cdot\|$ denotes the Frobenius norm. Moreover, assume that the gradient is bounded and
414 Lipschitz continuous. Then,*

$$\|W(t) - W_t^r\| \leq c_1 \varepsilon + c_2 \lambda + c_3 \vartheta / \lambda + c_4 \hat{\vartheta} / \lambda, \quad (9)$$

415 where the constants c_1, c_2, c_3 are independent of singular values of S^{-1} and S_v^{-1} .

416 *Proof.* We start by bounding the local error. That is, we assume that $W(t_0) = W_0^r$ and $\mathcal{V}(t_0) = \mathcal{V}_0^r$,
417 where \mathcal{V}_0^r is the momentum of the low-rank method. By definition of \hat{U} we have $(I - \hat{U} \hat{U}^\top) \mathcal{V}(t_0) = 0$
418 and thus

$$\|(I - \hat{U} \hat{U}^\top) \mathcal{V}(t)\| \leq \int_{t_0}^t \|(I - \hat{U} \hat{U}^\top) (\gamma \mathcal{V}(s) + P(W(s)) \nabla_W \mathcal{L}(W(s)))\| ds.$$

419 Using the boundedness of normal components and a Taylor expansion around t_0 gives for $s \in [t_0, t_1]$

$$\begin{aligned}P(W(s)) \nabla_W \mathcal{L}(W(s)) &= \nabla_W \mathcal{L}(W(s)) + O(\varepsilon) = \nabla_W \mathcal{L}(W(t_0)) + O(\lambda + \varepsilon) \\ &= P(W(t_0)) \nabla_W \mathcal{L}(W(t_0)) + O(\lambda + \varepsilon).\end{aligned} \quad (10)$$

420 Hence, with $\mathcal{V}(s) = \mathcal{V}(t_0) + O(\lambda + \varepsilon)$,

$$\begin{aligned}\|(I - \hat{U} \hat{U}^\top) \mathcal{V}(t)\| &\leq \lambda \|(I - \hat{U} \hat{U}^\top) (\gamma \mathcal{V}(t_0) + P(W(t_0)) \nabla_W \mathcal{L}(W(t_0)))\| + O(\lambda^2 + \lambda \varepsilon) \\ &= \lambda \|(I - \hat{U} \hat{U}^\top) \nabla_W \mathcal{L}(W(t_0)) V_0 V_0^\top\| + O(\lambda^2 + \lambda \varepsilon).\end{aligned}$$

By construction of \widehat{U} we have $0 = (I - \widehat{U}\widehat{U}^\top)\nabla_U \mathcal{L}(W(t_0)) = (I - \widehat{U}\widehat{U}^\top)\nabla_W \mathcal{L}(W(t_0))V_0$, hence

$$\|(I - \widehat{U}\widehat{U}^\top)\mathcal{V}(t)\| \leq O(\lambda^2 + \lambda\varepsilon).$$

From this, we directly conclude

$$\|(I - \widehat{U}\widehat{U}^\top)W(t_1)\| = \|(I - \widehat{U}\widehat{U}^\top)(W(t_0) + \int_{t_0}^{t_1} \mathcal{V}(s) ds)\| = O(\lambda^3 + \lambda^2\varepsilon).$$

An analogous derivation for the co-range gives

$$\begin{aligned} \|W(t_1) - \widehat{U}\widehat{U}^\top W(t_1)\widehat{V}\widehat{V}^\top\| &\leq \|(I - \widehat{U}\widehat{U}^\top)W(t_1)\| + \|W(t_1)(I - \widehat{V}\widehat{V}^\top)\| \\ &= O(\lambda^3 + \lambda^2\varepsilon). \end{aligned}$$

Next, we need to bound

$$\|\widehat{U}\widehat{U}^\top W(t_1)\widehat{V}\widehat{V}^\top - \widehat{U}\widehat{S}^1\widehat{V}^\top\| \leq \|\widehat{U}^\top W(t_1)\widehat{V} - \widehat{S}^1\|. \quad (11)$$

We note that from (10) we have with $W_0 := W(t_0)$ and $\mathcal{V}_0 := \mathcal{V}(t_0)$

$$\begin{aligned} \widehat{U}^\top W(t_1)\widehat{V} &= \widehat{U}^\top(W_0 + \lambda(1 - \gamma)\mathcal{V}_0 - \lambda^2 P(W_0)\nabla_W \mathcal{L}(W_0))\widehat{V} + O(\lambda^2 + \lambda\varepsilon) \\ &= \bar{S} - \lambda\gamma\bar{S}_v - \lambda\widehat{U}^\top \nabla_W \mathcal{L}(W_0)\widehat{V} + O(\lambda^2 + \lambda\varepsilon), \end{aligned}$$

where $\bar{S} = \widehat{U}^\top W_0 \widehat{V}$ and $\bar{S}_v = \widehat{U}^\top \mathcal{V}_0 \widehat{V}$. By definition of the S -update of Algorithm 2 we have

$$\widehat{S}^1 = \bar{S} + \lambda(1 - \gamma)\bar{S}_v - \lambda^2 \nabla_{\bar{S}} \mathcal{L}(\widehat{U}\bar{S}\widehat{V}^\top).$$

Thus, since $\nabla_{\bar{S}} \mathcal{L}(\widehat{U}\bar{S}\widehat{V}^\top) = U^\top \nabla_W \mathcal{L}(W_0)\widehat{V}$ we have $\|\widehat{U}^\top W(t_1)\widehat{V} - \widehat{S}^1\| = O(\lambda^2 + \lambda\varepsilon)$ and therefore the local error is bounded by

$$\begin{aligned} \|W(t_1) - W_1^\top\| &\leq \|W(t_1) - \widehat{U}\widehat{U}^\top W(t_1)\widehat{V}\widehat{V}^\top\| + \|\widehat{U}\widehat{U}^\top W(t_1)\widehat{V}\widehat{V}^\top - \widehat{U}\widehat{S}^1\widehat{V}^\top\| \\ &= O(\lambda^2 + \lambda\varepsilon). \end{aligned}$$

From the truncation tolerance ϑ , the bound on the truncation of \mathcal{V} , and the stability of the exact flow, we can obtain the desired error bound for the global error using Lady Windermere's fan. \square

We remark, that we can always ensure that condition $\|\widehat{U}_t \widehat{U}_t^\top \mathcal{V}_t \widehat{V}_t \widehat{V}_t^\top - U_t U_t^\top \mathcal{V}_t V_t V_t^\top\| \leq \widehat{\vartheta}$, is fulfilled for a user determined $\widehat{\vartheta}$, e.g. $\widehat{\vartheta} = \vartheta$, by increasing the new rank r_1 in the truncation step of Algorithm 3 if necessary.

9 Details to the numerical experiments of this work

9.1 UCM and Cifar Benchmarks

9.1.1 Network architecture and training details

In this paper, we use the pytorch implementation for neural network training. We take pretrained weights from the imagenet1k dataset as initialization, except for the long-term training study using ViT-small, which is randomly initialized. The data-loader randomly samples a batch for each batch-update which is the only source of randomness in our training setup. Below is an overview of the used network architectures

- VGG16 is a deep convolutional neural network architecture that consists of 16 layers, including 13 convolutional layers and 3 fully connected layers.
- VGG11 is a convolutional neural network architecture similar to VGG16 but with fewer layers, consisting of 11 layers: 8 convolutional layers and 3 fully connected layers. It follows the same design principle as VGG16, using small 3x3 convolution filters and 2x2 max-pooling layers.
- ViT32b is a Vision Transformer with 32x32 patch size, a deep learning architecture that leverages transformer models for image classification tasks.
- ViT-small is a compact vision transformer with patch size 8×8 , and an embedding dimension of 512. The model comprises six attention layers, each equipped with two heads, followed by a ResNet block and a dropout layer.

Table 3: Training hyperparameters for the UCM, Cifar10 and Cifar100 Benchmark. The first set hyperparameters apply to both DLRT and baseline training, and we train DLRT with the same hyperparameters as the full-rank baseline models. The second set of hyper-parameters is specific to DLRT. The DLRT hyperparameters are selected by an initial parameter sweep. We choose the DLRT truncation tolerance relative to the Frobenius norm of \hat{S} , i.e. $\vartheta = \tau \|\hat{S}\|_F$, as suggested in [22].

Hyperparameter	VGG16	VGG11	ViT-16b	ViT-small
Batch Size (UCM)	16	16	16	n/a
Batch Size (Cifar10, Cifar100)	128	128	128	256
Learning Rate	0.001	0.001	0.001	0.0001
Number of Epochs (UCM, Cifar10)	20	20	5	450
Number of Epochs (Cifar100)	30	30	20	n/a
L2 regularization	0	0	0.001	1e-2
Optimizer	AdamW	AdamW	AdamW	Adam
DLRT rel. truncation tolerance τ	0.1	0.05	0.08	0.05
Initial Rank	150	150	150	200

The full training setup is described in Table 3. We train DLRT with the same hyperparameters as the full-rank baseline models. It is known [21] that DLRT methods are robust w.r.t. common hyperparameters as learning rate, and batch-size, and initial rank. The truncation tolerance τ is chosen per an initial parameter study. These values are similar to default values reported in recent literature [23, 24, 26]. In general, there is a trade-off between target compression ratio and accuracy, as illustrated e.g. in [22] for matrix-valued and [26] for tensor-valued (CNN) layers.

9.1.2 UCM Data

The UC Merced (UCM) Land Use Dataset [31] is a standard benchmark in remote sensing and computer vision. It consists of 2,100 high-resolution aerial RGB images, each of size 256×256 pixels, organized into 21 land use classes with 100 images per class.

We normalize the training and validation data using channel-wise means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225]. Convolutional neural networks (CNNs) are applied directly to the original 256×256 image resolution. For the Vision Transformer (ViT), the input images are resized to 224×224 pixels within the data pipeline.

9.1.3 CIFAR-10 Data

The CIFAR-10 dataset comprises 60,000 RGB images of size 32×32 pixels, uniformly distributed across 10 object classes.

We apply standard data augmentation techniques to the training set, including random horizontal flipping followed by normalization with mean [0.4914, 0.4822, 0.4465] and standard deviation [0.2470, 0.2435, 0.2616]. The test set is only normalized. The same augmentation strategy is applied to CIFAR-100, using mean [0.5071, 0.4867, 0.4408] and standard deviation [0.2673, 0.2564, 0.2762].

CNNs are trained on the original 32×32 resolution, while ViT models receive images resized to 224×224 through the data pipeline. All adversarial attacks for this dataset are conducted on the resized inputs.

9.2 GLUE Benchmark

9.2.1 Dataset description

We present the benchmark overview in Table 4. We evaluate ALG against several recent fine-tuning methods on the General Language Understanding Evaluation (GLUE) benchmark [28]. GLUE is a standard benchmark comprising a diverse set of natural language understanding tasks that assess a model’s ability to comprehend and process human language. It provides a broad evaluation by including tasks covering various linguistic aspects such as entailment, sentiment, and semantic similarity. The benchmark comprises the following nine tasks:

Table 4: Summary of GLUE benchmark tasks

Corpus	Task	#Train	#Dev	#Test	#Label	Metrics
Single-Sentence Classification (GLUE)						
CoLA	Acceptability	8.5k	1k	1k	2	Matthews corr
SST2	Sentiment	67k	872	1.8k	2	Accuracy
Pairwise Text Classification (GLUE)						
MNLI	NLI	393k	20k	20k	3	Accuracy
RTE	NLI	2.5k	276	3k	2	Accuracy
QQP	Paraphrase	364k	40k	391k	2	F1
MRPC	Paraphrase	3.7k	408	1.7k	2	Accuracy
QNLI	QA/NLI	108k	5.7k	5.7k	2	Accuracy
Text Similarity (GLUE)						
STS-B	Similarity	7k	1.5k	1.4k	1	Pearson/Spearman cor

- **CoLA (Corpus of Linguistic Acceptability)**: Determines if a sentence is grammatically acceptable.
- **SST-2 (Stanford Sentiment Treebank)**: A binary sentiment classification task distinguishing between positive and negative sentiment.
- **MRPC (Microsoft Research Paraphrase Corpus)**: Identifies whether two given sentences are paraphrases.
- **STS-B (Semantic Textual Similarity Benchmark)**: Measures the semantic similarity of two sentences on a continuous scale from 1 to 5.
- **QQP (Quora Question Pairs)**: Assesses whether two questions are semantically equivalent.
- **QNLI (Question Natural Language Inference)**: Determines if a context sentence correctly answers a question.
- **RTE (Recognizing Textual Entailment)**: A binary entailment classification task.
- **Specific Focus**: MRPC (Microsoft Research Paraphrase Corpus)

The F1 score, used for evaluation, is computed from the precision P and recall R as follows. The precision P is defined as

$$P := \frac{P_T}{P_T + P_F}, \quad (12)$$

where P_T denotes the number of true positives and P_F the number of false positives. The recall R is given by

$$R := \frac{P_T}{P_T + N_F}, \quad (13)$$

where N_F represents the number of false negatives. The F1 score is then the harmonic mean of P and R :

$$F1 := \frac{2PR}{P + R}. \quad (14)$$

9.2.2 Reference implementations

Full Finetuning (FT): The standard approach in transfer learning, where the model is initialized with pre-trained weights and all parameters are updated via gradient descent.

Bitfit [32]: Finetuning where only the bias terms are updated while all other parameters remain fixed.

Adapter Tuning [10, 19]: Involves inserting two-layer adapter modules within transformer blocks. In [10], adapters are placed between the self-attention and feed-forward modules with a residual connection (denoted HAdapter). In [19], adapters are inserted after the feed-forward and layer normalization modules (denoted PAdapter), following the notation of [35].

LoRA [11]: Applies low-rank additive updates to selected weight matrices, modeled as

$$\mathbf{z} = \sigma \left(W_{\text{pt}} \mathbf{x} + \frac{\alpha}{r} AB^\top \mathbf{x} \right), \quad (15)$$

where $A, B \in \mathbb{R}^{n \times r}$. We apply LoRA to the attention matrices W_q, W_k, W_v , and the feed-forward matrices W_{f_1} and W_{f_2} . Learning rates and optimizers follow the setup in [35], Appendix D–F.

Results for FT, Bitfit, Adapter tuning, and LoRA in Table 2 are reproduced from [35]. The performance of DoRA, LoRA, LoRA+, and AdaLoRA is computed using the HuggingFace implementations of these adapters.

DoRA [17]: A low-rank adapter similar in structure to LoRA, but with normalized AB matrices and an additional magnitude parameter. Unlike LoRA, DoRA initializes the adapter with the pre-trained weights W_0 , rather than zero.

LoRA+ [6]: Differs from LoRA in the assignment of learning rates: separate learning rates are used for A and B , with a fixed ratio $\lambda_B/\lambda_A = 1.1$.

AdaLoRA [35]: Introduces adaptive low-rank updates to selected weight matrices:

$$\mathbf{z} = \sigma \left(W_{\text{pt}} \mathbf{x} + \frac{\alpha}{r} U S V^\top \mathbf{x} \right), \quad (16)$$

with frozen base weights $W_{\text{pt}} \in \mathbb{R}^{n \times n}$, rank- r adapters $U, V \in \mathbb{R}^{n \times r}$, and scaling matrix $S \in \mathbb{R}^{r \times r}$. The rank is determined using either SVD-based truncation or sensitivity analysis of the singular vectors. AdaLoRA is applied to W_q, W_k, W_v, W_{f_1} , and W_{f_2} with an orthogonality regularization coefficient $\gamma = 0.1$.

When comparing to AdaLoRA, we align the total parameter budget with LoRA by setting the final budget $b^{(T)}$ to 576, and initialize with $b^{(0)} = 1.5 \times b^{(T)}$.

We also compare AdaLoRA using budget schedules obtained via Algorithm 2, ensuring that $b^{(T)}$ approximately matches the parameter count of the final models trained using Algorithm 2.

GeoLoRA [21]: GeoLoRA integrates the projected gradient flow Equation (5) in a parallelizable single-step scheme, including a rank adaptive augmentation-truncation scheme as the proposed method. However, the method is only applicable for stochastic gradient descent, and not yet extended to momentum-based approaches. We use the hyperparameter choices reported in [21].

We use the implementation of [35] Appendix C] to compute the results for the presented reference methods. We set the exponential moving average parameters β_1 and β_2 of AdamW as their pytorch default value. We select the learning rates as denoted in Table 5 selected by an initial hyperparameter sweep.

We implement ALG as similar as possible to the reference models to achieve a fair comparison. That is, we add an adapter of the form $\mathbf{z} = \sigma(W_{\text{pt}} \mathbf{x} + U S V^\top \mathbf{x})$ to the key W_k , query W_q and value W_v matrices of all attention blocks, and to both feed-forward layers W_{f_1} and W_{f_2} . For each adapter, we employ Algorithm 2 to update the layer weights and ranks.

Table 5: Hyper-parameter setup for the GLUE benchmark, determined by an initial hyperparameter sweep.

Dataset	Learning Rate	Batch Size	# Epochs	τ	init. rank	Adapter dropout	weight decay
RTE	1.2×10^{-3}	32	20	0.075	10	0.01	0.01
QNLI	5×10^{-4}	64	5	0.05	10	0.2	0.01
MRPC	1×10^{-4}	64	5	0.05	10	0.15	0.05
QQP	1×10^{-4}	64	5	0.05	10	0.15	0.05
SST-2	1×10^{-4}	64	10	0.05	10	0.05	0.01
CoLA	5×10^{-4}	32	25	0.05	10	0.1	0.01
STS-B	1×10^{-3}	128	30	0.05	10	0.05	0.1

9.3 Computational hardware

All experiments in this paper are computed using workstation GPUs. Each training run used a single GPU. Specifically, we have used 5 NVIDIA RTX A6000, and 3 NVIDIA RTX 4090.

The estimated time for one experimental run depends mainly on the data-set size and neural network architecture. For training, generation of adversarial examples and validation testing we estimate 30 minutes on one GPU for one run.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist",**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We describe our contribution in the introduction section.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the underlying assumptions of the method in the analysis of the corresponding theorems.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We discuss the underlying assumptions of the method in the analysis of the corresponding theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: All experimental details are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We provide the open source code upon paper acceptance

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We provide the training details and hyperparameters in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: We provide error bars and report the mean and median over different initializations.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: The used compute resources are reported in the appendix

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The paper conforms, in every respect, with the NeurIPS Code of Ethics

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The impacts are discussed in the conclusion

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work is algorithmic and does not release special data or models.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Not needed for this work

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Not needed for this work

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Not needed for this work

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Not needed for this work

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.