A Detailed Proofs for XGBLoRA Lemmas

Lemma 4 (XGBLoRA Gradient Approximation) *The XGBLoRA update approximates the full gradient update with error:*

$$\|\nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}}\mathcal{L}(\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T})-\mathbf{A}^{(t)}\mathbf{B}^{(t)T}\|_F \leq \frac{C_1}{\sqrt{r}}+\frac{C_2}{\sqrt{M}}$$

where r is the LoRA rank, M is the number of minibatches, and C_1, C_2 are constants.

Proof 1 *1) Let $\mathbf{G}=\nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}}\mathcal{L}(\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T})$ be the true gradient.*

2) The XGBLoRA update $\mathbf{A}^{(t)}\mathbf{B}^{(t)T}$ can be seen as an approximation of $\mathbf{G}$.

3) Let $\mathbf{G}_r$ be the best rank-r approximation of $\mathbf{G}$. By the Eckart-Young-Mirsky theorem:

$$\|\mathbf{G}-\mathbf{G}_r\|_F \leq \frac{\|\mathbf{G}\|_*}{\sqrt{r}} \leq \frac{C_1}{\sqrt{r}}$$

where $\|\cdot\|_*$ is the nuclear norm and C_1 is a constant depending on the properties of $\mathcal{L}$.

4) The XGBLoRA update $\mathbf{A}^{(t)}\mathbf{B}^{(t)T}$ is computed using M minibatches. Let $\mathbf{G}_j$ be the gradient estimate from the j-th minibatch. Then:

$$\mathbf{A}^{(t)}\mathbf{B}^{(t)T} \approx \frac{1}{M}\sum_{j=1}^{M}\mathbf{G}_j$$

5) By the law of large numbers and assuming bounded variance of gradient estimates:

$$\|\frac{1}{M}\sum_{j=1}^{M}\mathbf{G}_j-\mathbf{G}\|_F \leq \frac{C_2}{\sqrt{M}}$$

where C_2 is a constant related to the gradient variance.

6) Combining these bounds using the triangle inequality:

$$\|\mathbf{G}-\mathbf{A}^{(t)}\mathbf{B}^{(t)T}\|_F \leq \|\mathbf{G}-\mathbf{G}_r\|_F+\|\mathbf{G}_r-\mathbf{A}^{(t)}\mathbf{B}^{(t)T}\|_F \leq \frac{C_1}{\sqrt{r}}+\frac{C_2}{\sqrt{M}}$$

This completes the proof.

Lemma 5 (Accumulated Update Bound) *For the XGBLoRA update process:*

$$\|\mathbf{A}^{(t)}\|_F \leq \eta_m MG \quad \textit{and} \quad \|\mathbf{B}^{(t)}\|_F \leq \eta_m MG$$

where G is an upper bound on $\|\nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}}\mathcal{L}(\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T})\|_F$.

Proof 2 *1) Recall the update rule for $\mathbf{A}^{(t)}$:*

$$\mathbf{A}^{(t)} \leftarrow \mathbf{A}^{(t)}-\eta_m\nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}}\mathcal{L}(\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T})\mathbf{B}^{(t)}$$

2) Taking the Frobenius norm and applying the triangle inequality:

$$\|\mathbf{A}^{(t)}\|_F \leq \|\mathbf{A}^{(t-1)}\|_F+\eta_m\|\nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}}\mathcal{L}(\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T})\|_F\|\mathbf{B}^{(t)}\|_F$$

3) Using the gradient bound $\|\nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}}\mathcal{L}(\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T})\|_F \leq G$:

$$\|\mathbf{A}^{(t)}\|_F \leq \|\mathbf{A}^{(t-1)}\|_F+\eta_m G\|\mathbf{B}^{(t)}\|_F$$

4) Applying this inequality recursively for all M minibatches, and noting that $\mathbf{A}^{(t)}$ is initialized to $\mathbf{0}$:

$$\|\mathbf{A}^{(t)}\|_F \leq \eta_m MG\|\mathbf{B}^{(t)}\|_F$$

5) Similarly for $\mathbf{B}^{(t)}$, we can derive:

$$\|\mathbf{B}^{(t)}\|_F \leq \eta_m MG \|\mathbf{A}^{(t)}\|_F$$

6) Combining these inequalities:

$$\|\mathbf{A}^{(t)}\|_F \leq \eta_m MG \quad \textit{and} \quad \|\mathbf{B}^{(t)}\|_F \leq \eta_m MG$$

This completes the proof.

Lemma 6 (Gradient Lipschitz Continuity) *For any two weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$:*

$$\|\nabla_{\mathbf{W}_1}\mathcal{L}(\mathbf{W}_1) - \nabla_{\mathbf{W}_2}\mathcal{L}(\mathbf{W}_2)\|_F \leq L'\|\mathbf{W}_1 - \mathbf{W}_2\|_F$$

where L is the Lipschitz constant of the gradient.

Proof 3 *1) This lemma is a standard assumption in optimization theory, often referred to as the smoothness condition.*

2) It can be derived from the assumption that the Hessian of $\mathcal{L}$ is bounded:

$$\|\nabla^2\mathcal{L}(\mathbf{W})\|_2 \leq L \quad \forall \mathbf{W}$$

where $\|\cdot\|_2$ denotes the spectral norm.

3) By the mean value theorem, there exists a $\mathbf{W}_t = t\mathbf{W}_1 + (1-t)\mathbf{W}_2$ for some $t \in [0,1]$ such that:

$$\nabla_{\mathbf{W}_1}\mathcal{L}(\mathbf{W}_1) - \nabla_{\mathbf{W}_2}\mathcal{L}(\mathbf{W}_2) = \nabla^2\mathcal{L}(\mathbf{W}_t)(\mathbf{W}_1 - \mathbf{W}_2)$$

4) Taking the Frobenius norm of both sides:

$$\|\nabla_{\mathbf{W}_1}\mathcal{L}(\mathbf{W}_1) - \nabla_{\mathbf{W}_2}\mathcal{L}(\mathbf{W}_2)\|_F = \|\nabla^2\mathcal{L}(\mathbf{W}_t)(\mathbf{W}_1 - \mathbf{W}_2)\|_F$$

5) Using the property that $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_F$:

$$\|\nabla^2\mathcal{L}(\mathbf{W}_t)(\mathbf{W}_1 - \mathbf{W}_2)\|_F \leq \|\nabla^2\mathcal{L}(\mathbf{W}_t)\|_2\|\mathbf{W}_1 - \mathbf{W}_2\|_F$$

6) Applying the bound on the Hessian:

$$\|\nabla^2\mathcal{L}(\mathbf{W}_t)\|_2\|\mathbf{W}_1 - \mathbf{W}_2\|_F \leq L\|\mathbf{W}_1 - \mathbf{W}_2\|_F$$

This completes the proof.

B Detailed Proof of XGBLoRA Convergence Theorem

Theorem 3 (XGBLoRA Convergence) *Under the XGBLoRA update process, assuming β-smoothness and μ-strong convexity of $\mathcal{L}$, after T iterations:*

$$\mathbb{E}[\mathcal{L}(\mathbf{W}^{(T)})] - \mathcal{L}^* \leq \frac{C_3}{\sqrt{T}} + \frac{C_4}{NT} + \epsilon(r)$$

where C_3 and C_4 are constants depending on β, μ, G, η_m, L, and $\epsilon(r) = \frac{C_5}{r}$ for some constant C_5.

Proof 4 *1) Let $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \mathbf{A}^{(t)}\mathbf{B}^{(t)T}$ be the update at iteration t.*

2) By the β-smoothness of $\mathcal{L}$:

$$\begin{aligned}\mathcal{L}(\mathbf{W}^{(t+1)}) &\leq \mathcal{L}(\mathbf{W}^{(t)}) + \langle\nabla\mathcal{L}(\mathbf{W}^{(t)}), \mathbf{A}^{(t)}\mathbf{B}^{(t)T}\rangle + \frac{\beta}{2}\|\mathbf{A}^{(t)}\mathbf{B}^{(t)T}\|_F^2 \\ &\leq \mathcal{L}(\mathbf{W}^{(t)}) + \langle\nabla\mathcal{L}(\mathbf{W}^{(t)}), \mathbf{A}^{(t)}\mathbf{B}^{(t)T}\rangle + \frac{\beta}{2}\|\mathbf{A}^{(t)}\|_F^2\|\mathbf{B}^{(t)}\|_F^2\end{aligned}$$

3) Using the XGBLoRA Gradient Approximation Lemma:

$$\mathbf{A}^{(t)}\mathbf{B}^{(t)T} = \nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}}\mathcal{L}(\mathbf{W}^{(t)} + \mathbf{A}^{(t)}\mathbf{B}^{(t)T}) + \mathbf{E}^{(t)}$$

where $\|\mathbf{E}^{(t)}\|_F \leq \frac{C_1}{\sqrt{r}} + \frac{C_2}{\sqrt{M}}$.

4) Substituting this into the inequality from step 2:

$$\mathcal{L}(\mathbf{W}^{(t+1)}) \leq \mathcal{L}(\mathbf{W}^{(t)}) + \langle \nabla \mathcal{L}(\mathbf{W}^{(t)}), \nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}} \mathcal{L}(\mathbf{W}^{(t)} + \mathbf{A}^{(t)}\mathbf{B}^{(t)T}) + \mathbf{E}^{(t)} \rangle + \frac{\beta}{2}\|\mathbf{A}^{(t)}\|_F^2 \|\mathbf{B}^{(t)}\|_F^2$$

5) Using the Gradient Lipschitz Continuity Lemma:

$$\|\nabla \mathcal{L}(\mathbf{W}^{(t)}) - \nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}} \mathcal{L}(\mathbf{W}^{(t)} + \mathbf{A}^{(t)}\mathbf{B}^{(t)T})\|_F \leq L' \|\mathbf{A}^{(t)}\mathbf{B}^{(t)T}\|_F$$

6) Applying Cauchy-Schwarz inequality and the bound from step 5:

$$\mathcal{L}(\mathbf{W}^{(t+1)}) \leq \mathcal{L}(\mathbf{W}^{(t)}) - \|\nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}} \mathcal{L}(\mathbf{W}^{(t)} + \mathbf{A}^{(t)}\mathbf{B}^{(t)T})\|_F^2 + L' \|\mathbf{A}^{(t)}\mathbf{B}^{(t)T}\|_F^2 + \|\nabla \mathcal{L}(\mathbf{W}^{(t)})\|_F \|\mathbf{E}^{(t)}\|_F + \frac{\beta}{2}\|\mathbf{A}^{(t)}\|_F^2 \|\mathbf{B}^{(t)}\|_F^2$$

7) Using the Accumulated Update Bound Lemma and the gradient bound G:

$$\mathcal{L}(\mathbf{W}^{(t+1)}) \leq \mathcal{L}(\mathbf{W}^{(t)}) - (1 - L\eta_m^2 M^2 G^2 - \frac{\beta}{2}\eta_m^2 M^2 G^2) \|\nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}} \mathcal{L}(\mathbf{W}^{(t)} + \mathbf{A}^{(t)}\mathbf{B}^{(t)T})\|_F^2 + G(\frac{C_1}{\sqrt{r}} + \frac{C_2}{\sqrt{M}})$$

8) By μ-strong convexity of $\mathcal{L}$:

$$\|\nabla_{\mathbf{W}^{(t)}+\mathbf{A}^{(t)}\mathbf{B}^{(t)T}} \mathcal{L}(\mathbf{W}^{(t)} + \mathbf{A}^{(t)}\mathbf{B}^{(t)T})\|_F^2 \geq 2\mu(\mathcal{L}(\mathbf{W}^{(t)} + \mathbf{A}^{(t)}\mathbf{B}^{(t)T}) - \mathcal{L}^*)$$

9) Substituting this into the inequality from step 7:

$$\mathcal{L}(\mathbf{W}^{(t+1)}) - \mathcal{L}^* \leq (1 - 2\mu\alpha)(\mathcal{L}(\mathbf{W}^{(t)}) - \mathcal{L}^*) + G(\frac{C_1}{\sqrt{r}} + \frac{C_2}{\sqrt{M}})$$

where $\alpha = 1 - L\eta_m^2 M^2 G^2 - \frac{\beta}{2}\eta_m^2 M^2 G^2$.

10) Taking expectation and applying this inequality recursively for T iterations:

$$\mathbb{E}[\mathcal{L}(\mathbf{W}^{(T)}) - \mathcal{L}^*] \leq (1 - 2\mu\alpha)^T (\mathcal{L}(\mathbf{W}^{(0)}) - \mathcal{L}^*) + \frac{G}{2\mu\alpha}(\frac{C_1}{\sqrt{r}} + \frac{C_2}{\sqrt{M}})$$

11) Using the inequality $(1-x)^T \leq \exp(-xT) \leq \frac{1}{xT}$ *for* $x \in (0,1)$*:*

$$\mathbb{E}[\mathcal{L}(\mathbf{W}^{(T)}) - \mathcal{L}^*] \leq \frac{C_3}{\sqrt{T}} + \frac{C_4}{M\sqrt{T}} + \frac{C_5}{r}$$

where $C_3 = \frac{(\mathcal{L}(\mathbf{W}^{(0)}) - \mathcal{L}^*)}{2\mu\alpha}$, $C_4 = \frac{GC_2}{2\mu\alpha}$, *and* $C_5 = \frac{GC_1}{2\mu\alpha}$.

This completes the proof.

C Detailed Proof of XGBLoRA Expressiveness Theorem

Theorem 4 (XGBLoRA Expressiveness) *Let f^* be any function in the original function class, and f_T be the function represented by the XGBLoRA-updated network after T iterations. Then:*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(f_T(\mathbf{x}) - f^*(\mathbf{x}))^2] \leq C_6(\frac{1}{r} + \frac{1}{M\sqrt{T}} + \frac{1}{\sqrt{T}})$$

where C_6 is a constant depending on the network architecture, the Lipschitz constants of the activation functions, and L.

Proof 5 *1) Let $\mathbf{W}^*$ be the weights that exactly represent f^* in the original function class.*

2) Define f_{opt} as the best function that can be represented by XGBLoRA *updates:*

$$f_{opt} = \arg\min_{f \in \mathcal{F}_{\text{XGBLoRA}}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(f(\mathbf{x}) - f^*(\mathbf{x}))^2]$$

where $\mathcal{F}_{\text{XGBLoRA}}$ is the class of functions representable by XGBLoRA *updates.*

3) We can decompose the error as:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(f_T(\mathbf{x}) - f^*(\mathbf{x}))^2] &\leq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(f_T(\mathbf{x}) - f_{opt}(\mathbf{x}))^2] + 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(f_{opt}(\mathbf{x}) - f^*(\mathbf{x}))^2] \\ &= 2E_1 + 2E_2\end{aligned}$$

4) For E_1, we can use the Convergence Theorem (Theorem 1):

$$E_1 \leq K_1(\frac{1}{\sqrt{T}} + \frac{1}{M\sqrt{T}})$$

where K_1 is a constant related to C_3 and C_4 from Theorem 1.

5) For E_2, we need to analyze how well XGBLoRA *updates can approximate $\mathbf{W}^*$. Let $\Delta\mathbf{W} = \mathbf{W}^* - \mathbf{W}^{(0)}$.*

6) We can approximate $\Delta\mathbf{W}$ with a sequence of low-rank updates:

$$\Delta\mathbf{W} \approx \sum_{t=1}^{T} \mathbf{A}^{(t)}(\mathbf{B}^{(t)})^T$$

7) By the properties of low-rank matrix approximation:

$$\|\Delta\mathbf{W} - \sum_{t=1}^{T} \mathbf{A}^{(t)}(\mathbf{B}^{(t)})^T\|_F \leq \frac{\|\Delta\mathbf{W}\|_*}{\sqrt{rT}}$$

where $\|\cdot\|_*$ denotes the nuclear norm.

8) Assuming the network function is Lipschitz continuous with respect to its weights with Lipschitz constant L_f:

$$E_2 \leq L_f^2\|\Delta\mathbf{W} - \sum_{t=1}^{T} \mathbf{A}^{(t)}(\mathbf{B}^{(t)})^T\|_F^2 \leq \frac{L_f^2\|\Delta\mathbf{W}\|_*^2}{rT}$$

9) Combining the bounds for E_1 and E_2:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[(f_T(\mathbf{x}) - f^*(\mathbf{x}))^2] &\leq 2K_1(\frac{1}{\sqrt{T}} + \frac{1}{M\sqrt{T}}) + \frac{2L_f^2\|\Delta\mathbf{W}\|_*^2}{rT} \\ &\leq C_6(\frac{1}{r} + \frac{1}{M\sqrt{T}} + \frac{1}{\sqrt{T}})\end{aligned}$$

where $C_6 = \max(2K_1, 2L_f^2\|\Delta\mathbf{W}\|_*^2)$.

This completes the proof.

D Broader Impact

The proposed XGBLoRA framework has the potential to bring about significant positive societal impacts by democratizing access to state-of-the-art language technologies. By enabling efficient and effective fine-tuning of large language models, XGBLoRA can empower researchers and practitioners with limited computational resources to leverage the power of pre-trained models for a wide range of downstream tasks. This can foster innovation and accelerate progress in various domains, such as healthcare, education, and social sciences, where natural language understanding and generation

Table 7: Details of GLUE dataset.

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

can be applied to improve decision-making, personalize learning experiences, and analyze large-scale social data. However, it is crucial to acknowledge and mitigate potential negative societal impacts associated with the widespread adoption of language models. Fine-tuned models may perpetuate biases present in the pre-training data, leading to unfair or discriminatory outcomes if not carefully audited and corrected. Additionally, the efficiency of `XGBLoRA` may lower the barrier to developing and deploying language models, potentially enabling malicious actors to create and disseminate harmful content at scale. To address these concerns, it is important to develop and adhere to ethical guidelines for the responsible development and deployment of language models, ensuring transparency, accountability, and fairness. Researchers and practitioners should also actively engage in public discourse to raise awareness about the benefits and risks of language technologies and collaborate with policymakers to develop appropriate governance frameworks. By proactively addressing these challenges, we can harness the potential of efficient fine-tuning techniques like `XGBLoRA` to create positive societal impact while mitigating the risks and negative consequences.

E LIMITATIONS

One limitation of our current approach is that our theoretical analysis is based on linear models, which may influence the generalizability of our findings to more complex, non-linear systems. Additionally, the assumptions made in our theoretical framework may not hold in certain real-world scenarios, potentially limiting the applicability of our method in such cases. Future work will focus on extending our theory to encompass more generalized forms, allowing for a broader range of applications and improved robustness to model misspecification.