LoRA Training Provably Converges to a Low-Rank Global Minimum or It Fails Loudly (But it Probably Won't Fail)

Junsu Kim

Seoul National University, Math

Jaeyeon Kim

Harvard, CS

Ernest K. Ryu

UCLA, Math

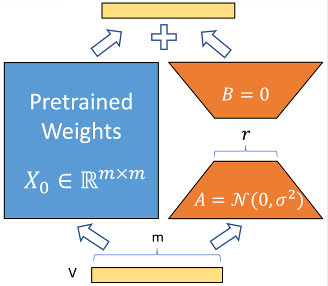
Junsu Kim is applying for Ph.D. programs this fall!

International Conference on Machine Learning
Oral Presentation, 2025

July 17, 2025

LoRA background

Low-Rank Adaptation (LoRA) fine-tunes large pre-trained foundation models by introducing low-rank updates to the attention layers.


Pretrained Weights
$X_0 \in \mathbb{R}^{m\times m}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
m
V

Given a linear layer mapping

$$V \mapsto VX_0$$

with X_0, LoRA introduces the rank-r update

$$V \mapsto V(X_0 + AB^\intercal)$$

with $A \in \mathbb{R}^{m\times r}$ and $B \in \mathbb{R}^{m\times r}$. The $X_0 \in \mathbb{R}^{m\times m}$ weights are frozen (not trained) while A and B are trained.

A is initialized randomly while B is zero-initialized.
So $AB^\intercal = 0$ at initialization.

$r = 16$ is a common choice for the rank r.

E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, *ICLR*, 2022.

LoRA for transformers

In a transformer, it is common practice to place LoRA on the linear layers for QKV and position-wise FFN.

Denote the pre-trained weights as X_0 and the fine-tuned updates as $X_\square$.

Full fine-tuning:

$$\underset{X_\square}{\mathsf{minimize}} \quad \hat{\mathcal{L}}(X_\square) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_{X_0+X_\square}(x_i), y_i).$$

LoRA fine-tuning:

$$\underset{A,B}{\mathsf{minimize}} \quad \hat{\mathcal{L}}(AB^\intercal) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_{X_0+AB^\intercal}(x_i), y_i).$$

Outline

Background and prior work

LoRA Training Provably Converges to a Low-Rank Global Minimum or It Fails Loudly

Weight decay on LoRA is nuclear norm regularization

LoRA training often uses weight decay separately on the LoRA factors. Can be interpreted as solving

$$\underset{A,B}{\text{minimize}} \quad \hat{\mathcal{L}}(AB^\intercal) + \tfrac{\lambda}{2}\|A\|_F^2 + \tfrac{\lambda}{2}\|B\|_F^2,$$

with regularization parameter $\lambda \geq 0$. This is equivalent to

$$\underset{X_\square,\, \mathrm{rank} X_\square \leq r}{\text{minimize}} \quad \hat{\mathcal{L}}_\lambda(X_\square) \triangleq \hat{\mathcal{L}}(X_\square) + \lambda \|X_\square\|_*,$$

where $X_\square = AB^\intercal$ and $\|\cdot\|_*$ is the nuclear norm (sum of singular values).

Insight: Weight decay induces nuclear norm regularization, which, in turn, induces low-rank updates.

B. Recht, M. Fazel, and P. A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, *SIAM Review*, 2010.

The NTK assumption

If the first-order Taylor approximation holds throughout training

$$f_{X_0+X_\square}(x) \approx f_{X_0}(x) + \langle \nabla f_{X_0}(x), X_\square \rangle$$

we say training stays within the NTK regime.

This approximation is justified empirically when prompt-based fine-tuning is used.

S. Malladi, A. Wettig, D. Yu, D. Chen, and S. Arora, A kernel-based view of language model fine-tuning, *ICML*, 2023.

Background: Strict saddles vs. SOSP

U is a (first-order) *stationary* point if

$$\nabla \hat{L}(U) = \mathbf{0}.$$

U is a *second-order stationary point* (SOSP) if

$$\nabla \hat{L}(U) = \mathbf{0}, \qquad \nabla^2 \hat{L}(U)[V, V] \geq 0,$$

for any direction $V \in \mathbb{R}^{m \times n}$. (Hessian has no negative eigenvalues.)

U is *strict saddle* if it is a first- but not second-order stationary point.

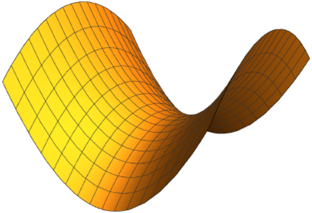


Figure: A strict saddle

Background: SGD avoids strict saddles

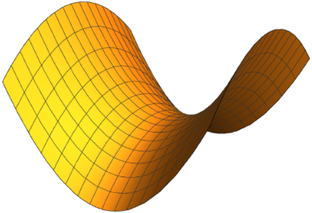


Figure: A strict saddle

Stochastic gradient descent (SGD) does not converge strict saddle points. SGD only converges to SOSP.

In general, however, an SOSP can be non-global local minima (spurious local minima). In our setup, all SOSPs are global minima, so SGD converges to global minima.

R. Ge, F. Huang, C. Jin, and Y. Yuan, Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition, *COLT*, 2015.

J. D. Lee, M. Simchowitz, M. I. Jordan, and Benjamin Recht, Gradient descent only converges to minimizers, *COLT*, 2016.

LoRA in the NTK regime has no spurious local minima

Theorem

Assume $\frac{r(r+1)}{2} > N$. (So $r \gtrsim \sqrt{N}$.) Consider the linearized neural network and consider a training loss with a small random perturbation. Then, all SOSPs are global minimizers with probability 1.

Theorem applies when the fine-tuning data size N is not too large: $N \sim 1000$ and $r \sim 30$.

Generically, LoRA training has no spurious local minima!

The training loss has saddle points, but SGD won't converge to them. SGD converges to an SOSP, which is a global minimum.

U. Jang, J. D. Lee, and E. K. Ryu, LoRA training in the NTK regime has no spurious local minima, ICML Oral, 2024.

Outline

Background and prior work

LoRA Training Provably Converges to a Low-Rank Global Minimum or It Fails Loudly

Assumptions in prior work

Prior theoretical works analyzing LoRA rely on some strong assumptions:

- Small N.
- Linearization, i.e., NTK. (Confirmed to hold in some but not all practical LoRA fine-tuning setups.)
- Other highly simplified setups.

In this work, we analyze LoRA under much relaxed assumptions and obtain qualitatively new of conclusions.

Assumption: Existence of a low-rank minimizer

We assume there exists a rank $r_\star$ global minimizer $X_\star$ of the full fine-tuning loss $\widehat{\mathcal{L}}_\lambda(X_0 + X_\square)$ (without any linearization) and that our LoRA module uses rank $r > r_\star$.

This is a strong assumption, but it (approximately) holds in most practical setups.

I don't know why many fine-tuning tasks admit low-rank updates. This work proves that if there is a low-rank update, then LoRA finds it.

Assumption: Restricted strong convexity and restricted smoothness

Deep learning objectives are typically neither strongly convex nor have small smoothness constants. However, they do satisfy *restricted* strong convexity and smoothness in many practical fine-tuning scenarios.

f is (α, r, D)-*restricted strongly convex* about $X_\star$ if

$$\langle \nabla f(X) - \nabla f(X_\star), X - X_\star \rangle \geq \alpha \|X - X_\star\|_F^2.$$

for any $X \in \mathbb{R}^{m \times n}$ such that $\|X - X_\star\|_F \leq D$ and $\operatorname{rank}(X) \leq r$.

f is (β, r, D)-*restricted smooth* about $X_\star$ if

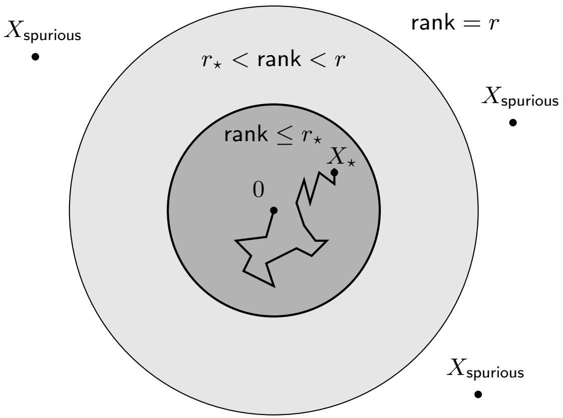
$$\nabla^2 f(X)[UX + XV, UX + XV] \leq \beta \|UX + XV\|_F^2$$

for any $[X \in \mathbb{R}^{m \times n}$ such that $\|X - X_\star\|_F \leq D$ and $\operatorname{rank}(X) \leq r]$, $[U \in \mathbb{R}^{m \times m}$ such that $\|U\|_F = \|V\|_F = 1]$ and $\operatorname{rank}(U) = 1]$, and $[V \in \mathbb{R}^{n \times n}$ such that $\|V\|_F = 1$ and $\operatorname{rank}(U) = \operatorname{rank}(V) = 1]$.

Weaker assumption than the linearization assumption of prior work.
(C.f. Discussion at the end of §3.1 of paper.)

These assumptions are more reasonable since (i) it only needs to hold locally and (ii) it only needs to hold for deviations of small rank.

Illustration of main result



Spurious local minima X_{spurious} may exist, but they have high rank and large magnitude. Since LoRA training starts at $X_\square = AB^\intercal = 0$ and since weight decay regularizes the rank (nuclear norm) of $X_\square$, training likely converges to the global minimum $X_\star$.

LoRA training converges to a global min. or fails loudly

Theorem

Let (A, B) be a SOSP of $\widehat{\mathcal{L}}_{\lambda}^{\mathrm{lora}}$ with $X_{\square} = AB^{\intercal}$ and $\|X_{\square} - X_{\star}\|_F \leq D$. Then,

(i) *If $\sigma_r(X_{\square}) \leq \frac{2\alpha}{\beta}\sigma_{r_\star}(X_{\square})$, then $X_{\square}$ is a global minimizer.*

(ii) *If $\sigma_r(X_{\square}) > \frac{2\alpha}{\beta}\sigma_{r_\star}(X_{\square})$, then $X_{\square}$ is a spurious solution, and further $X_{\square}$ has large magnitude with*

$$\|X_{\square}\|_F \geq \sqrt{\frac{\sum_{s=r_\star+1}^{r} \sigma_s^2(X_{\square})}{1 - \frac{2\alpha\sigma_{r_\star}}{\beta\sigma_r}}} - \|X_\star\|_F.$$

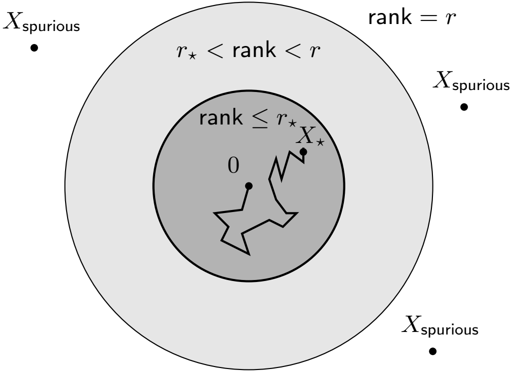
If LoRA fine-tuning does converge to a spurious solution, its high rank and large magnitude would be noticeable, and generalization will be poor. In this sense, we describe this mode of failure to be failing loudly.

In programming, “failing loudly” refers to coding practices that cause immediate, obvious failures (crashes or explicit errors) rather than quietly continuing with incorrect behavior. Loud failures make debugging easier as they are detected immediately.

LoRA training probably won't fail; it probably won't converge to spurious local minima

But we argue that this failure is unlikely due the following implicit biases:

- Zero-initialization ($X_\square = AB^\intercal = 0$ at initialization) biases the optimization towards minima with smaller magnitude.
- Weight decay (applied to the A and B factors separately) implicitly biases the optimization towards low-rank matrices.



Experiments: Verifying RSC and RSM

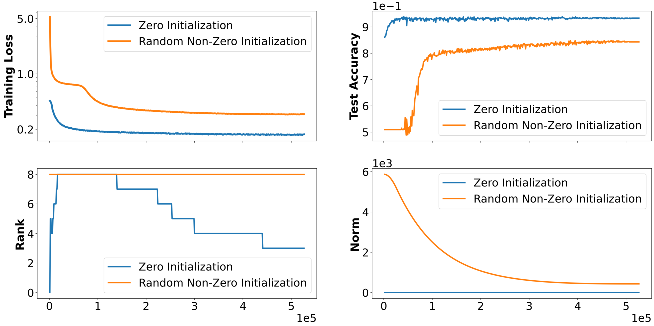
We estimate α and β with $r = 8, 16, 32, 64$, $D = 5$, and $\lambda = 0.01$.

Rank	8	16	32	64
β/α	8.0249	18.7032	320.82	N/A
α	0.0061	0.0029	0.0002	-0.0445
β	0.0492	0.0539	0.0726	0.3371

Findings: Assumption $\alpha > 0$ and $\beta < \infty$ are plausible when r is small.

This also suggests that reduced memory footprint is not the only benefit of using small r; the α, β-values that determine the loss landscape also become more favorable with small r.

Experiments: Validating main theorem

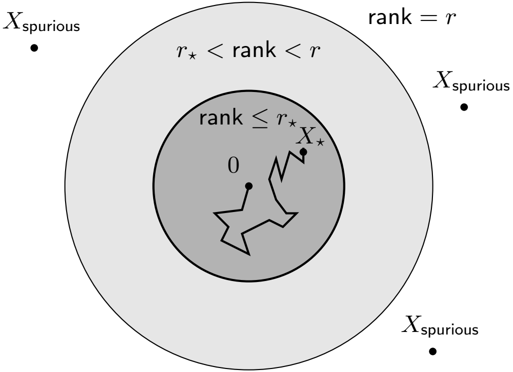


Findings:

- With zero-initialization, LoRA training converging to global minima.
- With random non-zero initialization, LoRA training converges to spurious local minima.

Conclusion

Using low-rank matrix-sensing machinery, we proved a new type of landscape result for LoRA.


X_{spurious}
rank $= r$
$r_\star <$ rank $< r$
X_{spurious}
rank $\leq r_\star$
$X_\star$
0
X_{spurious}