Low-Rank Adaptation Secretly Imitates Differentially Private SGD

Saber Malekmohammadi [1] **Golnoosh Farnadi** [1 2]

Abstract

We look at low-rank adaptation methods, e.g., LoRA, from the lens of data privacy. We show theoretically that the low-rank adaptation used in LoRA is equivalent to fine-tuning adapters with noisy batch gradients - just like what DPSGD algorithm does. We also quantify the variance of the induced noise as a decreasing function of adaptation rank. By establishing a Berry-Esseen type bound on the total variation distance between distribution of the induced noise and a Gaussian distribution with the same variance, we show that the dynamics of low-rank adaptation is very close to that when performing DPSGD w.r.t the adapters. Consequently, low-rank adaptation provides robustness to membership inference attacks.

1. Introduction

Stochastic Gradient Descent (SGD) is the power engine of training deep neural networks, which updates parameters of a model by using batch gradient descent. Modern deep learning models, e.g., GPT-3 (Brown et al., 2020) and Stable Diffusion (Rombach et al., 2022) have a large number of parameters, which induces a large space complexity for their fine-tuning with SGD. LoRA (Hu et al., 2021) updates only some of the parameters, called *adapters*, by restricting their updates to be a low-rank matrix. This low-rank restriction considerably reduces the number of trainable parameters, at the cost of limiting the optimization space of the adapters. As another parameter-efficient training technique, ReLoRA (Lialin et al., 2023) utilizes low-rank updates to train high-rank networks to mitigate the low-rank constraint of LoRA. Similarly, the work in (Hao et al., 2024) identifies that the dynamics of LoRA can be approximated by a random matrix projection. Based on this, the work proposes FLoRA to achieve high-rank updates by resampling the random projection matrices, which enables achieving higher utility while enjoying the sublinear space complexity of LoRA.

On the other hand, from the lens of data privacy, fine-tuning data often happens to be privacy-sensitive. DPSGD (Abadi et al., 2016) fine-tunes the existing parameters in a pre-trained model by using noisy clipped batch gradients. However, it induces a high space complexity (Abadi et al., 2016). As a remedy, DPSGD can be used to fine-tune only the adapters in a pre-trained model (Liu et al., 2024).

In this work, we draw a connection between low-rank adaptation and DPSGD w.r.t the adapters. We show that low-rank adaptation of the adapter parameters is equivalent to fine-tuning them with noisy batch gradients, which is very close to what DPSGD does for fine-tuning adapters privately. We also quantify the variance of the noise injected into the batch gradients as a decreasing function of the adaptation rank. Furthermore, we show that the distribution of the injected noise is very close to a Gaussian distribution with the same variance. Our derivations shows that low-rank adaptation and DPSGD w.r.t the adapters are very close to each other in terms of their mechanisms. The highlights of our contributions are as follows:

- We show that low-rank adaptation of adapters is equivalent to fine-tuning them with noisy batch gradients (Equation (3)).
- We quantify the variance of the noise injected into the batch gradients w.r.t the adapters, and show that the noise approaches a Gaussian distribution as the adaptation layer's input dimension increases (Lemma 3.1).
- We bound the total variation distance between the distribution of the injected noise and a pure Gaussian noise with the same variance. The bound decreases as the input dimension of the adaptation layer and the adaptation rank increase (Lemma 3.2).
- Based on the above findings, we show that the dynamics of low-rank adaptation is very close to that of fine-tuning the adapters with DPSGD (Section 4).
- Finally, we show experimentally that low-rank adaptation provides robustness to memebrship inference atatcks w.r.t the fine-tuning dataset (Section 5), confirming our theoretical findings.

[1]Mila - Quebec AI Institute [2]School of Computer Science, McGill University, Montreal, Canada. Correspondence to: Saber Malekmohammadi <saber.malekmohammadi@uwaterloo.ca>.

Published at ICML 2025 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Vancouver, Canada. PMLR 267, 2025.

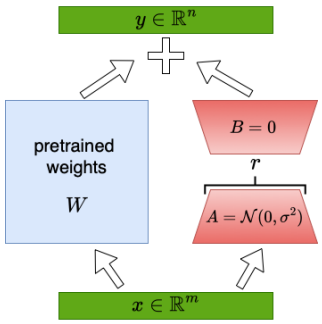

Figure 1. Low-rank decomposition of LoRA in an adaptation layer.

2. Dynamics of low-rank adaptation

We start by studying the dynamics of low-rank adaptation. In order to update a pre-trained adapter weight $W \in \mathbb{R}^{n\times m}$ located in a layer of a pre-trained model, LoRA incorporates low-rank decomposition matrices $B \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{r\times m}$, where $r \ll \min\{n, m\}$, and performs the following forward pass in an adaptation layer (see Figure 1):

$$y = (W + BA)x = Wx + BAx, \tag{1}$$

where $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ are the input and the pre-activation output of the current adaptation layer, respectively. For instance, there are 12 self-attention layers in GPT-2 model and when fine-tuning it with LoRA, W can be considered to be the query layer in each of those 12 layers, which has dimension $(m, n) = (768, 2304)$. For each adaptation layer, it is common to initialize B with an all-zero matrix and sample A from the normal distribution $\mathcal{N}(0, \frac{1}{r})$. We will use this intialization of A for deriving our theoretical results. When back-propagating, LoRA calculates the gradients w.r.t only A and B, which can be found as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial A} &= \frac{\partial BA}{\partial A} \cdot \frac{\partial \mathcal{L}}{\partial BA} = B^\top \cdot \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial y}{\partial BA} = B^\top(\nabla_W \mathcal{L}),\\ \frac{\partial \mathcal{L}}{\partial B} &= \frac{\partial \mathcal{L}}{\partial BA} \cdot \frac{\partial BA}{\partial B} = \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial y}{\partial BA} \cdot A^\top = (\nabla_W \mathcal{L})A^\top.\end{aligned} \tag{2}$$

Hence, $\frac{\partial \mathcal{L}}{\partial A} \in \mathbb{R}^{r\times m}$ and $\frac{\partial \mathcal{L}}{\partial B} \in \mathbb{R}^{n\times r}$. In fact, LoRA down-projects the batch gradient $\nabla_W \mathcal{L}$ from $\mathbb{R}^{n\times m}$ to a lower dimension, and updates the matrices A and B with the resulting projections of $\nabla_W \mathcal{L}$ (Hao et al., 2024).

Having understood the connection between low-rank adaptation in LoRA and random projection, in the next section, we show that adaptation of the matrices A and B in Equation (2) is equivalent to "fine-tuning" adapter W with a noisy version of batch gradient $\nabla_{W^t}\mathcal{L}^t$ at every time step t. This is our first step towards establishing a comparison framework between the dynamics of low-rank adaptation and that of DPSGD w.r.t the dapter W.

3. Random noise injection by low-rank adaptation

Let A^0, B^0 and W^0 denote the initial/pre-trained values of the matrices A, B and the adapter W. A recent work (Hao et al., 2024) showed that during low-rank adaptation with LoRA, it is mostly the matrix B that gets adapted and matrix A is almost frozen. Hence, LoRA can be approximated by freezing A at A^0. For simplicity of our analysis, we keep using this approximation of LoRA. In that case, as we have shown in Appendix C, performing T SGD steps w.r.t B^t is equivalent to updating the forward-pass parameter W^0 in Equation (1) to the following forward pass parameter:

$$W^0 - \eta \sum_{t=0}^{T-1} \Bigg[\underbrace{\nabla_{W^t}\mathcal{L}^t}_{\text{batch gradient w.r.t } W^t} + \underbrace{\nabla_{W^t}\mathcal{L}^t(A^{0\top}A^0 - \mathbb{I}_m)}_{\text{noise} \in \mathbb{R}^{n\times m}} \Bigg], \tag{3}$$

where $W^t = W^0 + B^tA^0$, i.e. the equivalent forward pass parameter at time step t. The second term in the sum represents a noise term introduced by the low-rank adaptation. Therefore, *"low-rank adaptation" of pretrained adapter W^0 (by introducing low-rank matrices A and B) is equivalent to "fine-tuning" W^0 with noisy batch gradients.*

We are now particularly interested in the behavior of this noise term, which is added to each batch gradient $\nabla_{W^t}\mathcal{L}^t$ in every step t. Recall that the entries of A^0 were sampled from $\mathcal{N}(0, \frac{1}{r})$ (see Figure 1), and that each of the r columns of $A^{0\top}$ is an m-dimensional Gaussian random variable. Consequently, $A^{0\top}A^0$ follows a Wishart distribution with r degrees of freedom (Bhattacharya & Burman, 2016), which is the multivariate generalization of the chi-squared distribution. Therefore, for any $q \in \mathbb{R}^{1\times m}$, $q \cdot (A^{0\top}A^0 - \mathbb{I}_m)$ is a weighted sum of multiple chi-squared random variables, which implies that the result follows a Gaussian distribution approximately, according to the Central Limit Theorem (CLT) (Bhattacharya et al., 2016). Therefore, we first state and prove the following lemma concerning the noise term in Equation (3).

Lemma 3.1. *Let $A \in \mathbb{R}^{r\times m}$ be a matrix with i.i.d entries sampled from $\mathcal{N}(0, \frac{1}{r})$. Given a fixed $q \in \mathbb{R}^{1\times m}$, the distributions of the elements of $q\cdot(A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$ approach the Gaussian distribution $\mathcal{N}(0, \frac{\|q\|^2}{r})$, as m increases.*

Table 1. Membership inference attack on GPT-2. LoRA (with both frozen and trainable A) shows robustness to the attacks in terms of AUC and TPR of MIA by mimicking DPSGD and incorporating noisy batch gradients w.r.t adapters during adaptation.

Dataset	**Enron**				**Pubmed**				**PTB**			
Evaluation Criterion	**Utility**	**MIA Success Metrics**			**Utility**	**MIA Success Metrics**			**Utility**	**MIA Success Metrics**		
Metric	PPL↓	AUC↓	@FPR 10%↓	@FPR 1%↓	PPL↓	AUC↓	@FPR 10%↓	@FPR 1%↓	PPL↓	AUC↓	@FPR 10%↓	@FPR 1%↓
Full fine-tuning	18.49	0.876	66.31	13.02	16.64	0.929	80.67	6.48	27.70	0.963	90.54	65.84
LoRA ($r = 16$)	19.66	0.757	40.56	8.08	18.82	0.816	50.43	2.68	31.18	0.961	89.97	59.88
LoRA (frozen A^0**,** $r = 16$)	21.90	0.614	16.76	4.49	20.85	0.546	14.93	2.37	37.93	0.952	82.23	55.58
DPSGD on adapters ($\epsilon = 20$)	27.74	0.541	13.62	1.19	27.16	0.532	12.21	1.78	44.53	0.542	55.21	9.87

We also prove the following lemma to show the closeness of the distribution of the elements of $q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$ to $\mathcal{N}(0, \frac{\|q\|^2}{r})$ for limited values of m.

Lemma 3.2. *Let $A \in \mathbb{R}^{r\times m}$ be a matrix with i.i.d entries sampled from $\mathcal{N}(0, \frac{1}{r})$. Given a fixed $q \in \mathbb{R}^{1\times m}$ with $0 < c \leq |q_i| \leq C$, let $u = q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$. Let u_i be the i-th element of u. There exists $z \sim \mathcal{N}(0, \frac{\|q\|^2}{r})$ such that*

$$Pr\{u_i \neq z\} \in \mathcal{O}\left(\frac{1}{\sqrt{mr}}\right). \tag{4}$$

The lemma above means that each element of $q\cdot(A^\top A-\mathbb{I}_m)$ follows a mixture of distributions: $\mathcal{N}(0, \frac{\|q\|^2}{r})$ with weight w_g and another distribution M - which we dont know - with weight $(1-w_g) \in \mathcal{O}(\frac{1}{\sqrt{mr}})$. For instance, when finetuning GPT-2 model with LoRA with $r = 16$, $1/\sqrt{mr} \approx 9\times10^{-3}$, showing the distribution of the injected noise is indeed very close to a Gaussian distribution. The larger mr, the closer the mixture distribution gets to $\mathcal{N}(0, \|q\|^2/r)$.

4. Comparing Low Rank Adaptation and DPSGD

Based on our understandings from Lemma 3.2, low rank adaptation (with rank r) of adapter $W \in \mathbb{R}^{n\times m}$ at time step t is equivalent to fine-tuning it with the noisy stochastic batch gradient $\tilde{\nabla}_{W^t}\mathcal{L}^t = \nabla_{W^t}\mathcal{L}^t + N^t$, where $N^t \in \mathbb{R}^{n\times m}$ is a noise-term with Gaussian-like distribution: $Pr\{N^t_{i,j} \neq z^t_i\} \in \mathcal{O}(1/\sqrt{mr})$, where $z^t_i \sim \mathcal{N}(0, \frac{\|[\nabla_{W^t}\mathcal{L}^t]_{i,:}\|^2}{r})$. In other words, *low-rank adaptation injects noise into each row of the batch gradient* $\nabla_{W^t}\mathcal{L}^t$, and the standard deviation of the noise added to the elements of the row i is proportional to the ℓ_2 norm of row i. Also, as mr grows (i.e., the input dimension of the adaptation layer or the adaptation rank increase), the distribution of noise element $N^t_{i,j}$ gets closer to $\mathcal{N}(0, \frac{\|[\nabla_{W^t}\mathcal{L}^t]_{i,:}\|^2}{r})$.

The operation above is very similar to what DPSGD (Abadi et al., 2016) does when it is applied to W (see Appendix B for a background about DPSGD). At the t-th gradient update step on a current adapter parameter W^t, DPSGD computes the following noisy clipped batch gradient on a batch $\mathcal{B}^t$:

$$\tilde{\nabla}_{W^t}\mathcal{L}^t = \underbrace{\left(\frac{1}{b}\sum_{i\in\mathcal{B}^t}\bar{\nabla}_{W^t}\mathcal{L}^t_i\right)}_{\text{batch gradient w.r.t } W^t} + \underbrace{\mathcal{N}(0, \frac{(c^t)^2 z^2}{b^2})}_{\text{noise} \in \mathbb{R}^{n\times m}}, \tag{5}$$

where $\bar{\nabla}_{W^t}\mathcal{L}^t_i$ is the clipped sample gradient computed on sample i with clipping threshold c^t. If $c^t \geq \|\nabla_{W^t}\mathcal{L}^t_i\|$, then the clipping will be ineffective and can be ignored. This is very similar to the noisy batch gradients used in Equation (3). LoRA adds random noise to different rows of the average batch gradient according to their norms: rows with larger norms experience a larger noise. Also, the injected noise to each element follows a Gaussian distribution with a high probability (Lemma 3.2). Accordingly, we expect low-rank adaptation to provide informal robustness to privacy attacks against the fine-tuning data, just like DPSGD.

5. Experiments

In this section, we perform membership inference attacks (MIA) on some fine-tuned models to show how low-rank adaptation indeed provides robustness against them. We use model perplexity as a measure of utility. We also use TPR (true positive rate) and AUC to measure the MIA success rate. See the short paragraph in Appendix A for complete description of the experimental setup.

Experimental results: In Table 1, we have reported the results obtained from evaluating MIA effectiveness on the GPT-2 model and different fine-tuning algorithms. As observed, full fine-tuning all parameters achieves the best utility (PPL) on all datasets with the cost of a considerable vulnerability to MIA. In contrast, choosing a subset of parameters in the existing 12 layers of the model (the adapters) and adapting them with LoRA improves robustness to MIA considerably. As observed, LoRA achieves a considerably lower AUC and TPR@FPR compared to full fine-tuning at the cost of a slight drop in utility (larger perplexity), thanks to the low-rank adaptation mechanism incorporating noisy batch gradients w.r.t to the adapter parameters. Also, we can observe that LoRA with a frozen A^0 enhances this robustness even further. For comparison, we have included the results for fine-tuning the same adapter parameters in the existing 12 layers with DPSGD. As expected, fine-tuning with

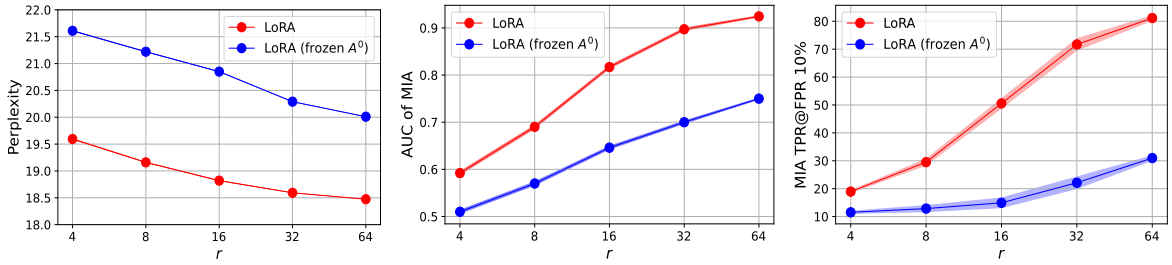

Figure 2. The ffect of adaptation rank r on the model fine-tuned on Pubmed averaged over three different data splits (random seeds). **Left:** The perplexity (utility) of the fine-tuned model decreases (increases) as r increases. **Middle:** The AUC of MIA attacks increases as r increases, showing less robustness to attacks for lager r. **Right:** The success rate of MIA increases as r increases.

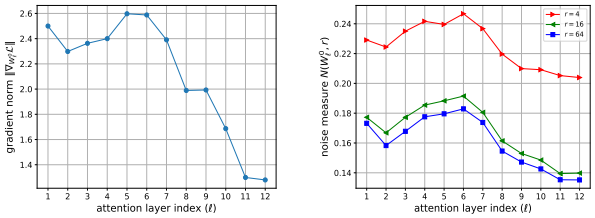

Figure 3. Experimental results obtained on Pubmed dataset showing that, unlike DPSGD, LoRA adds noise to batch gradient w.r.t the adapter parameters *non-uniformly*. **Left:** the norm of the batch gradient w.r.t the adapters in different layers of the GPT-2 model. **Right:** The noise measure $N(W_\ell^0, r)$ evaluated at different layers (indexed by ℓ) of the model for different ranks r.

DPSGD also improves the robustness to MIA compared to full fine-tuning. However, it leads to a more severe drop in utility, which can be attributed to the fact that DPSGD adds noise uniformly with the same variance to all the elements in batch gradients. In contrast, LoRA tends to add smaller noise to the rows in batch gradients that have a smaller norm. It is also noteworthy that, in all our experiments, running DPSGD needed more than twice as much runtime as LoRA needed, due to its higher space complexity. Figure 4 in appendix summarizes the results in Table 1 more clearly.

According to Lemma 3.1 and Lemma 3.2, we expect LoRA to show more robustness to MIA as r decreases. As observed in Figure 2, the perplexity (and robustness to MIA) of the fine-tuned model increases as r decreases. In this sense, the adaptation rank r in LoRA behaves very similar to the privacy parameter ϵ of DPSGD in adjusting the robustness of the fine-tuned model to MIA.

Now, we introduce a measure to quantify the amount of noise introduced by low-rank adaptation. Let W_ℓ^0 denote the pretrained adapter parameter at the layer ℓ of the model. Also, let W_ℓ^1 denote the adapter parameter that is obtained for the layer ℓ after one SGD update with step size η w.r.t W_ℓ^0. In that case, the forward pass of the adaptation layer ℓ will be updated to $y = W_\ell^1 x$ (Figure 1). Similarly, let A_ℓ^1 and B_ℓ^1 denote the parameters obtained from low-rank adaptation of W_ℓ^0 (with rank r) after one SGD update with the same step size η w.r.t A_ℓ^0 and B_ℓ^0 (when matrix A_ℓ^0 is frozen, we will have $A_\ell^1 = A_\ell^0$). In that case, according to Equation (12), the forward pass of the adaptation layer ℓ will be updated to $y = (W_\ell^0 + B_\ell^1 A_\ell^1)x$. According to our understandings that low-rank adaptation of W_ℓ^0 is the same as fine-tuning it with noisy batch gradients, the noise amount introduced by low-rank adaptation to the adapter in layer ℓ after one gradient step can be measured by:

$$N(W_\ell^0, r) = \left\| W_\ell^1 - (W_\ell^0 + B_\ell^1 A_\ell^1) \right\|. \tag{6}$$

We have used the above noise measure in Figure 3 to confirm our theoretical findings that LoRA adds more noise to the adapter parameters that have a batch gradient with larger norm. We clearly observe that the adapter $W_6^0 \in \mathbb{R}^{768\times 2304}$ in the sixth attention layer of GPT-2, which has a largest $\|\nabla_{W_6^0}\mathcal{L}\|$ (left figure), experiences the largest noise in one gradient step of LoRA (tight figure). Also, as seen in the right figure, the amount of noise injected to different layers decreases uniformly as the rank r increases. We already observed in Figure 2 that larger r results in more vulnerability of the adapted model to privacy attacks, altogether confirming our theoretical findings in Lemma 3.1 and Section 4. Therefore, low-rank adaptation indeed provides robustness against MIA attacks by imitating the mechanism of DPSGD.

6. Conclusion

We uncovered a close similarity between the dynamics of low-rank adaptation and DPSGD. We showed that low-rank adaptation is equivalent to injecting random noise into the batch gradients w.r.t the adapter parameters non-uniformly. Accordingly, it provides robustness against privacy attacks w.r.t the fine-tuning data and this robustness increases as adaptation rank decreases.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016. URL http://dx.doi.org/10.1145/2976749.2978318.

Bhattacharya, P. and Burman, P. Multivariate analysis. In *Theory and Methods of Statistics*, pp. 383–429. Academic Press, 2016. ISBN 978-0-12-802440-9. URL https://www.sciencedirect.com/science/article/pii/B9780128024409000126.

Bhattacharya, R., Lin, L., and Patrangenaru, V. *A Course in Mathematical Statistics and Large Sample Theory*. Springer Texts in Statistics. Springer New York, 2016. URL https://books.google.ca/books?id=AgTWDAAAQBAJ.

Billingsley, P. *Probability and Measure*. John Wiley & Sons, Inc., 1995. ISBN 0471007102. URL https://www.colorado.edu/amath/sites/default/files/attached-files/billingsley.pdf.

Bobkov, S. G., Chistyakov, G. P., and Götze, F. Berry–esseen bounds in the entropic central limit theorem. *Probability Theory and Related Fields*, pp. 435–478, 2011. URL https://link.springer.com/content/pdf/10.1007/s00440-013-0510-3.pdf.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles, 2022. URL https://arxiv.org/abs/2112.03570.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. A discourse-aware attention model for abstractive summarization of long documents, 2018. URL https://arxiv.org/abs/1804.05685.

Devroye, L., Mehrabian, A., and Reddad, T. The total variation distance between high-dimensional gaussians with the same mean, 2023. URL https://arxiv.org/abs/1810.08693.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 2006. URL https://doi.org/10.1007/11761679_29.

Feller, W. *An introduction to probability theory and its applications*. John Wiley & Sons, Inc., 1971. URL https://www.google.ca/books/edition/An_Introduction_to_Probability_Theory_an/rxadEAAAQBAJ?hl=en&gbpv=0.

Gaunt, R. E. Absolute moments of the variance-gamma distribution, 2024. URL https://arxiv.org/abs/2404.13709.

gustav Esseen, C. Fourier analysis of distribution functions. a mathematical study of the laplace-gaussian law. *Acta Mathematica*, 77, 1945. URL https://link.springer.com/article/10.1007/BF02392223.

Hao, Y., Cao, Y., and Mou, L. Flora: Low-rank adapters are secretly gradient compressors. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/hao24a.html.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Klimt, B. and Yang, Y. The enron corpus: A new dataset for email classi(cid:12)cation research. 2004. URL https://api.semanticscholar.org/CorpusID:13451873.

Levin, D. A., Peres, Y., and Wilmer, E. L. *Markov chains and mixing times*. American Mathematical Society, 2008. URL https://www.cs.cmu.edu/~15859n/RelatedWork/MarkovChains-MixingTimes.pdf.

Lialin, V., Shivagunde, N., Muckatira, S., and Rumshisky, A. Relora: High-rank training through low-rank updates, 2023. URL https://arxiv.org/abs/2307.05695.

Liu, R., Wang, T., Cao, Y., and Xiong, L. Precurious: How innocent pre-trained language models turn into privacy traps, 2024. URL https://arxiv.org/abs/2403.09562.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, pp. 313–330, 1993. URL https://aclanthology.org/J93-2004/.

Mireshghallah, F., Uniyal, A., Wang, T., Evans, D., and Berg-Kirkpatrick, T. Memorization in nlp fine-tuning methods, 2022. URL https://arxiv.org/abs/2205.12506.

Mood, A. and Franklin, A. *Introduction to the Theory of Statistics*. McGraw-Hill, Inc., 1974. ISBN 0070428646. URL https://sistemas.fciencias.unam.mx/~misraim/Mood.pdf.

Petrov, V. V. *Sums of Independent Random Variables*. De Gruyter, 1975. URL https://doi.org/10.1515/9783112573006.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf.

Sablayrolles, A., Douze, M., Ollivier, Y., Schmid, C., and Jégou, H. White-box vs black-box: Bayes optimal strategies for membership inference, 2019. URL https://arxiv.org/abs/1908.11229.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., and Shokri, R. Enhanced membership inference attacks against machine learning models, 2022. URL https://arxiv.org/abs/2111.09679.

Appendix for *Low-Rank Adaptation Secretly Imitates Differentially Private SGD*

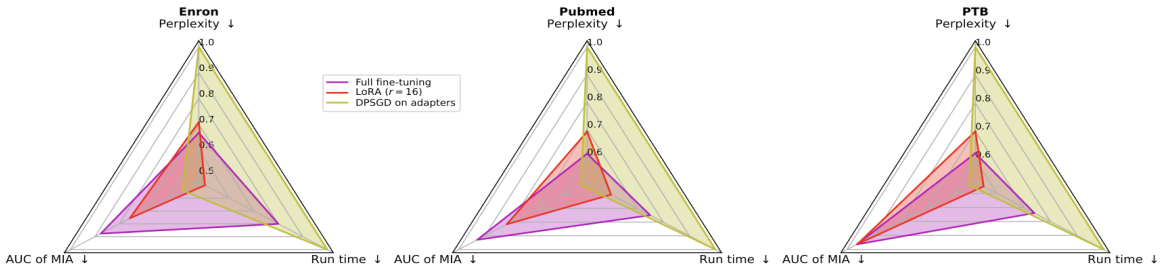

Figure 4. The trade-off existing between utility, privacy and computational/space complexity (run time) for different adaptation algorithms. LoRA provides robustness against privacy attacks to fine-tuning data without much adverse effect on utility or additional computational/space complexity.

A. Experimental setup

Datasets: We run our experiments on three fine-tuning datasets with confidential properties, including Penn Treebank (PTB)(Marcus et al., 1993), Enron (Klimt & Yang, 2004) and Pubmed (Cohan et al., 2018). We split each dataset into three parts and use them as train, validation and auxilliary datasets.

Models: We perform our experiments with GPT-2 (12-layer, 125M parameters, vocab size 50257) on two Nvidia A100 GPUs. We apply LoRA (Hu et al., 2021) with default $r = 16$ to the query parameters in the 12 attention layers of GPT-2, as in (Liu et al., 2024). For comparison, we also consider a baseline algorithm that fine-tunes the same set of adapter parameters with DPSGD without using any low-rank decomposition matrices. We also consider fine-tuning *all* parameters as another baseline. We use $E = 20$ epochs for fine-tuning with the default learning rates {1e-5, 5e-5, 5e-4} (with a linear scheduler) for full fine-tuning, fine-tuning adapters with DPSGD and low-rank adaptation with LoRA, respectively.

Membership Inference Attacks: We use a calibrated membership score for membership inference (Carlini et al., 2022; Mireshghallah et al., 2022; Sablayrolles et al., 2019; Ye et al., 2022). More specifically, we train a reference model θ_{ref} on the auxiliary dataset and use the following signal for classifying a sample x as a member/non-member of the fine-tuning dataset:

$$I_\theta(x) := \mathbb{I}[\mathcal{L}(x;\theta) - \mathcal{L}(x;\theta_{ref}) < \gamma]. \tag{7}$$

The threshold γ is set to the highest value for which the false positive rate over all samples will be lower than α. We consider $\alpha \in \{10\%, 1\%\}$ and report the corresponding true positive rates (TPR).

Metrics: To measure utility, we use the perplexity on the validation set (PPL). For measuring the effectiveness of MIA, we use AUC and TPR@FPR α for $\alpha \in \{10\%, 1\%\}$.

B. Background

In this section, we provide some background about differential privacy and DPSGD algorithm, which are used in the paper. In machine learning, differential privacy is often used to provide formal and rigorous data privacy guarantees w.r.t to training data. The following definition of differential privacy is one of the most commonly used definitions:

Definition B.1 ((ϵ, δ)-DP, (Dwork et al., 2006))**.** A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-DP if for any two adjacent inputs d, $d' \in \mathcal{D}$, which differ only by a single record, and for any measurable subset of outputs $\mathcal{S} \subseteq \mathcal{R}$ it holds that

$$\Pr[\mathcal{M}(d) \in \mathcal{S}] \leq e^\epsilon \Pr[\mathcal{M}(d') \in \mathcal{S}] + \delta.$$

Gaussian mechanism, which randomizes the output of a non-private computation f on a dataset d as $\mathbf{G}_\sigma f(d) \triangleq f(d) + \mathcal{N}(\mathbf{0}, \sigma^2)$, provides (ϵ, δ)-DP. The variance of the noise, σ^2, is calibrated to the sensitivity of f, i.e., the maximum amount

of change in its output (measured in ℓ_2 norm) on two neighboring datasets d and d'. Gaussian mechanism has been used in DPSGD algorithm (Abadi et al., 2016) for private ML by randomizing intermediate data-dependent computations, e.g., batch gradients. More specifically, at the t-th gradient update step on a current model $\boldsymbol{\theta} \in \mathbb{R}^p$, DPSGD computes the following noisy batch gradient:

$$\tilde{g}(\boldsymbol{\theta}) = \frac{1}{b}\Big[\Big(\sum_{i\in\mathcal{B}^t} \bar{g}_i(\boldsymbol{\theta})\Big) + \mathcal{N}(\mathbf{0}, \sigma^2_{\texttt{DP}}\mathbb{I}_p)\Big], \tag{8}$$

where $\bar{g}_i(\boldsymbol{\theta}) = \texttt{clip}(g_i, c)$, g_i is the sample gradient computed on the sample i in the batch $\mathcal{B}^t$ and c is a clipping threshold. For a given vector $\mathbf{v}$, $\texttt{clip}(\mathbf{v}, c) = \min\{\|\mathbf{v}\|, c\} \cdot \frac{\mathbf{v}}{\|\mathbf{v}\|}$. Also, $\sigma_{\texttt{DP}} = c \cdot z$, where z is the noise scale that should be used by DPSGD in order to achieve $(\epsilon, \delta)-$DP (for desired values of ϵ and δ) with respect to train data at the end of training. It can be computed by using a privacy accountant, e.g., the moments accountant (Abadi et al., 2016).

C. Derivation of Equation (3)

Let A^0, B^0 and W^0 denote the initial/pre-trained values of the matrices A, B and the adapter W (coming from the pretrained model). We restate the following thorem from (Hao et al., 2024) without restating its proof:

Theorem C.1 ((Hao et al., 2024)). *Let LoRA update matrices A and B with SGD as:*

$$A^{t+1} \leftarrow A^t - \eta\frac{\partial \mathcal{L}}{\partial A^t} = A^t - \eta B^{t\top}(\nabla_{W^t}\mathcal{L}^t), \tag{9}$$

$$B^{t+1} \leftarrow B^t - \eta\frac{\partial \mathcal{L}}{\partial B^t} = B^t - \eta(\nabla_{W^t}\mathcal{L}^t)A^{t\top}, \tag{10}$$

where η is the learning rate. We assume $\|\sum_{t=0}^{T}\nabla_{W^t}\mathcal{L}^t\|_F \le L$ for every T, implying that the model stays within a finite Euclidean ball from W^0. In this case, the dynamics of A^t and B^t are given by

$$A^t = A^0 + \eta A^0 f_A(t), \quad B^t = \eta f_B(t) A^{0\top}, \tag{11}$$

where the forms of $f_A(t) \in \mathbb{R}^{m\times m}$ and $f_B(t) \in \mathbb{R}^{n\times m}$ are expressed in the proof. ***In particular,*** *$\|f_A(t)\|_2 \le \frac{\eta L^2\left(1-(\eta^2L^2)^t\right)}{1-\eta^2L^2}$* ***for every*** *$t$.*

In the equation above, $W^t = W^0 + B^tA^t$, i.e. the equivalent forward pass parameter at time step t. Let us denote the total changes of A and B after T gradient steps as ΔA and ΔB, respectively. Then, the forward pass in Equation (1) will get updated to:

$$\big(W^0 + (B^0 + \Delta B)(A^0 + \Delta A)\big)x = \big(W^0 + \Delta BA^0 + \Delta B\Delta A\big)x, \tag{12}$$

where we have substituted the initialization $B^0 = \mathbf{0} \in \mathbb{R}^{n\times r}$. From Equation (11) and after substituting the values of ΔA and ΔB after T gradient steps into the above equation, we get to the following updated forward pass parameter for the adaptation layer:

$$W^0 + \Delta BA^0 + \Delta B\Delta A = W^0 + \eta f_B(T)A^{0\top}A^0 + \eta^2 f_B(T)A^{0\top}A^0 f_A(T). \tag{13}$$

Also, from Theorem C.1, we have $\|f_A(T)\|_2 \le \|f_A(T)\|_F \le \frac{\eta L^2\left(1-(\eta^2L^2)^T\right)}{1-\eta^2L^2}$. Hence, if $\eta \ll 1/L$, we have $\lim_{T\to\infty}\eta\|f_A(T)\|_2 = \lim_{T\to\infty}\frac{(\eta L)^2\left(1-(\eta L)^{(2T)}\right)}{1-(\eta L)^2} \ll 1$. Therefore, the last term on the right side is significantly

smaller than the second term $\eta f_B(T)A^{0\top}A^0$. Hence, the second term is the dominant term in the update to the forward pass parameter W^0. Therefore, we can closely approximate LoRA by freezing A to its initial value A^0 and adapting only the matrix B. In this case:

$$W^0 + \Delta BA^0 + \Delta B\Delta A = W^0 + \Delta BA^0 = W^0 + \eta \tilde{f}_B(T)A^{0\top}A^0, \tag{14}$$

where $\tilde{f}_B(0) = \mathbf{0}$ and $\tilde{f}_B(t+1) = \tilde{f}_B(t) - \nabla_{W^t}\mathcal{L}^t$. Equivalently, $\tilde{f}_B(T) = -\sum_{t=0}^{T-1} \nabla_{W^t}\mathcal{L}^t$. Substituting this into the equation above, we get:

$$W^0 + \Delta BA^0 + \Delta B\Delta A = W^0 + \Delta BA^0 = W^0 - \eta \sum_{t=0}^{T-1} \left[(\nabla_{W^t}\mathcal{L}^t)A^{0\top}A^0\right], \tag{15}$$

where the last term shows the exact change in the forward pass parameter W^0 after performing T rounds of SGD on the matrix B. **The term $\nabla_{W^t}\mathcal{L}^t$ in the sum is the batch gradient that would be obtained if we had "fine-tuned" the adapter W using SGD**. Therefore, low rank adaptation with LoRA can be viewed as performing a random projection of stochastic batch gradient $\nabla_{W^t}\mathcal{L}^t$ in every step t with matrix $A^{0\top}$ and projecting it back with matrix A^0. Finally, Equation (15) can be directly rewritten as Equation (3).

D. Useful Theorems

In this section, we mention some theorems, which we will use in our proofs.

Theorem D.1 (Chi-Squared distribution: (Mood & Franklin, 1974), Section 4.3, Theorem 7)**.** *If the random variables X_i, $i = 1, \ldots, k$, are normally and independently distributed with means μ_i and variances σ_i^2, then*

$$U = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \tag{16}$$

has a chi-squared distribution with k degrees of freedom: $U \sim \mathcal{X}_k^2$. Also, $\mathbb{E}[U] = k$ and $\texttt{Var}[U] = 2k$.

The theorem above states that sum of the squares of k standard normal random variables is a chi-squared distribution with k degrees of freedom.

Lemma D.2 (Raw moment of Chi-Squared distribution)**.** *Suppose $X \sim \mathcal{X}_k^2$. Then, the m-th raw moment of X can be found as follows;*

$$\mathbb{E}[X^m] = \prod_{i=0}^{m-1} (k + 2i) \tag{17}$$

Proof. From the definition of Chi-Squared distribution with r degrees of reddom, U has the following probability density function:

$$f_X(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \tag{18}$$

Therefore, we have:

$$\begin{aligned}\mathbb{E}[X^m] &= \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\int_0^{+\infty} x^{\frac{k}{2}+m-1}e^{-\frac{x}{2}}dx = \frac{2}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\int_0^{+\infty}(2u)^{\frac{k}{2}+m-1}e^{-u}du \\ &= \frac{2^{\frac{k}{2}+m-1+1}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}\int_0^{+\infty} u^{\frac{k}{2}+m-1}e^{-u}du = \frac{2^m}{\Gamma(\frac{k}{2})}\Gamma(\frac{k}{2}+m) = \frac{2^m\Gamma(\frac{k}{2})}{\Gamma(\frac{k}{2})}\prod_{i=0}^{m-1}(\frac{k}{2}+i) \\ &= \prod_{i=0}^{m-1}(k+2i). \end{aligned} \tag{19}$$

Note that the fifth equality directly results from the property of gamma function that for $z > 0$, $\Gamma(1+z) = z\Gamma(z)$. □

Theorem D.3 (Classical Central Limit Theorem: (Billingsley, 1995), Theorem 27.1)**.** *Suppose that $\{X_i\}_{i=1}^n$, is an independent sequence of random variables having the same distribution with mean μ and positive variance σ^2. Define $S_n = \sum_{i=1}^n X_i$ as their sum. Let Z_n be defined by*

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}. \tag{20}$$

Then, the distribution of Z_n approaches standard normal distribution as n approaches infinity.

The theorem above states that S_n is approximately, or asymptotically, distributed as a normal distribution with mean $n\mu$ and variance $n\sigma^2$.

The next theorem is about the Lindeberg's condition, which is a sufficient (and under certain conditions also a necessary condition) for the Central Limit Theorem (CLT) to hold for a sequence of independent random variables $\{X_i\}_{i=1}^n$. Unlike the classical CLT stated above, which requires the sequence of random variables to have a finite variance and be both independent and identically distributed (*i.i.d*), Lindeberg's CLT only requires the sequence of random variables to have finite variance, be independent and also satisfy the Lindeberg's condition. The following states the theorem.

Theorem D.4 (Lindeberg and Lyapounov Theorem: (Billingsley, 1995), Theorem 27.2)**.** *Suppose $X_1, \ldots, X_n$ are n independent random variables with $\mathbb{E}[X_i] = \mu_i$ and $\mathtt{Var}[X_i] = \sigma_i^2 > 0$. Define $S_n = \sum_{i=1}^n X_i$ and let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. Also assume the following condition holds for all $\epsilon > 0$:*

$$\textit{Lindeberg's condition:} \quad \lim_{n\to\infty}\sum_{i=1}^{n}\frac{1}{s_n^2}\int_{|x-\mu_i|\geq\epsilon s_n}(x-\mu_i)^2 P_{X_i}(x)dx = 0. \tag{21}$$

where P_{X_i} is the pdf of variable X_i. Assuming $Z_n = \frac{S_n - \sum_{i=1}^n \mu_i}{s_n}$, the distribution of Z_n approaches standard normal distribution as n approaches infinity.

The theorem above states that, given that Lindeberg's condition is satisfied, S_n is approximately, or asymptotically, distributed as a normal distribution with mean $\sum_{i=1}^n \mu_i$ and variance s_n^2, even if the sequence of variables are not identically distributed.

E. Proof of Lemma 3.1

Using the theorems above, we are now able to prove Lemma 3.1.

Proof. From the theorem's assumption, we know that elements of A are sampled from $\mathcal{N}(0, \frac{1}{r})$. Therefore, we can rewrite the product $q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$ as the following product:

$$q \cdot (\frac{A^\top A}{r} - \mathbb{I}_m) \in \mathbb{R}^{1\times m} \tag{22}$$

where *the elements of A are now from standard normal distribution*. Let $a_{i,j}$ denote the element in i-th row and j-th column of this new A. Therefore, for all i and j, $a_{i,j}$ has distribution $\mathcal{N}(0,1)$. Let $B = \frac{A^\top A}{r} - \mathbb{I}_m$. Also, let $A_{i,:}$ and $A_{:,j}$ denote the i-th row and j-th column of the new A, respectively. We have:

$$B_{i,i} = \frac{1}{r}[A^\top A]_{i,i} - 1 = \frac{1}{r}A_{:,i}^\top A_{:,i} - 1 = \frac{1}{r}\|A_{:,i}\|_2^2 - 1 = (\frac{1}{r}\sum_{l=1}^{r} a_{l,i}^2) - 1 \tag{23}$$

From Equation (22), we know that $a_{l,i}$ is from standard normal distribution. Hence, $a_{l,i}^2$ is a chi-squared with 1 degree of freedom: $a_{l,i}^2 \sim \mathcal{X}_1^2$. Therefore, $\sum_{l=1}^{r} a_{l,i}^2$, which is the sum of r independent chi-squared variables with 1 degree of freedom, is a chi-squared with r degrees of freedom: $\sum_{l=1}^{r} a_{l,i}^2 \sim \mathcal{X}_r^2$ (see Theorem D.1). Therefore, for $i \in \{1,\dots,m\}$, we have:

$$\begin{aligned}
&\mathbb{E}[B_{i,i}] = \mathbb{E}\big[\frac{\sum_{l=1}^{r} a_{l,i}^2}{r}\big] - 1 = \frac{r}{r} - 1 = 0,\\
&\texttt{Var}[B_{i,i}] = \texttt{Var}[\frac{\sum_{l=1}^{r} a_{l,i}^2}{r}] = \frac{\texttt{Var}(\mathcal{X}_r^2)}{r^2} = \frac{2r}{r^2} = \frac{2}{r}.
\end{aligned} \tag{24}$$

Similarly, we find the mean and variance of the non-diagonal elements $B_{i,j} (i \neq j)$ of B. We have:

$$B_{i,j} = \frac{1}{r}[A^\top A]_{i,j} = \frac{1}{r}A_{:,i}^\top A_{:,j} = \frac{1}{r}\sum_{l=1}^{r} a_{l,i}a_{l,j}, \tag{25}$$

where $a_{l,i}$ and $a_{l,j}$ are independent and standard normal. Therefore, $a_{l,i} + a_{l,j} \sim \mathcal{N}(0,2)$. Similarly, $a_{l,i} - a_{l,j} \sim \mathcal{N}(0,2)$. So we can rewrite $a_{l,i}a_{l,j}$ as:

$$a_{l,i}a_{l,j} = \frac{1}{4}(a_{l,i} + a_{l,j})^2 - \frac{1}{4}(a_{l,i} - a_{l,j})^2 = \frac{1}{2}z_1^2 - \frac{1}{2}z_2^2, \tag{26}$$

where z_1 and z_2 are from standard normal. Therefore, $a_{l,i}a_{l,j} = \frac{\nu_1 - \nu_2}{2}$, where $\nu_1, \nu_2 \sim \mathcal{X}_1^2$. Also, $a_{l,i} + a_{l,j}$ and $a_{l,i} - a_{l,j}$ are independent variables. Hence, z_1 and z_2 are independent, and likewise ν_1 and ν_2 are independent. We conclude that:

$$a_{l,i}a_{l,j} = \frac{1}{2}(\nu_1 - \nu_2), \tag{27}$$

where $\nu_1, \nu_2 \sim \mathcal{X}_1^2$, and are independent.

Now, lets assume $\nu_1, \nu_2 \sim \mathcal{X}_k^2$ (a more general case), and let $M_{\nu_1}(t) = \mathbb{E}[e^{t\nu_1}]$ be the moment generating function (MGF) of ν_1. In this case, we know that $M_{\nu_1}(t) = M_{\nu_2}(t) = (1-2t)^{-\frac{k}{2}}$ (MGF of $\mathcal{X}_k^2$). Hence, $M_{\nu_1-\nu_2}(t) = M_{\nu_1}(t) \cdot M_{\nu_2}(-t) = (1-4t^2)^{-\frac{k}{2}} = \big(\frac{\frac{1}{4}}{\frac{1}{4}-t^2}\big)^{\frac{k}{2}}$, which is the MGF of a symmetric about origin variance-gamma distribution with parameters $\lambda = \frac{k}{2}, \alpha = \frac{1}{2}, \beta = 0, \mu = 0, \gamma = \frac{1}{2}$. Therefore, when $\nu_1, \nu_2 \sim \mathcal{X}_k^2$, then $\nu_1 - \nu_2$ has this distribution, which has mean $\mu + 2\beta\lambda/\gamma^2 = 0$ and variance $2\lambda(1 + 2\beta^2/\gamma^2)/\gamma^2 = 4k$.

In Equation (27), we had $k = 1$, as we had $\nu_1, \nu_2 \sim \mathcal{X}_1^2$. Hence, based on the discussion above, we have:

$$\mathbb{E}[a_{l,i}a_{l,j}] = 0 \tag{28}$$

$$\texttt{Var}[a_{l,i}a_{l,j}] = \frac{1}{4}\texttt{Var}[\nu_1 - \nu_2] = \frac{4k}{4} = 1 \tag{29}$$

Consequently, based on Equation (25) and from the results above, we can compute the mean and variance of the non-diagonal elements of B ($i \neq j$):

$$\mathbb{E}[B_{i,j}] = \mathbb{E}\big[\frac{\sum_{l=1}^{r} a_{l,i}a_{l,j}}{r}\big] = \frac{\sum_{l=1}^{r} \mathbb{E}[a_{l,i}a_{l,j}]}{r} = 0,$$
$$\texttt{Var}[B_{i,j}] = \texttt{Var}[\frac{\sum_{l=1}^{r} a_{l,i}a_{l,j}}{r}] = \frac{\sum_{l=1}^{r} \texttt{Var}[a_{l,i}a_{l,j}]}{r^2} = \frac{r}{r^2} = \frac{1}{r}. \tag{30}$$

So far, we have computed the mean and variance of each entry in $B = \frac{A^\top A}{r} - \mathbb{I}_m \in \mathbb{R}^{m\times m}$ in Equation (24) and Equation (30). Now, for a given $q \in \mathbb{R}^{1\times m}$, we have:

$$q \cdot B = \sum_{l=1}^{m} q_l B_{l,:}, \tag{31}$$

where $B_{l,:}$ is row l of B. Let u_i denote the i-th element of $q \cdot B$. Hence, for each element u_i ($i \in \{1, \dots, m\}$), we have:

$$\mathbb{E}[u_i] = \mathbb{E}\Big[\sum_{l=1}^{m} q_l B_{l,i}\Big] = \sum_{l=1}^{m} q_l \mathbb{E}[B_{l,i}] = 0,$$
$$\texttt{Var}[u_i] = \texttt{Var}\Big[\sum_{l=1}^{m} q_l B_{l,i}\Big] = \sum_{l=1}^{m} q_l^2 \texttt{Var}[B_{l,i}] = q_i^2 \texttt{Var}[B_{i,i}] + \sum_{l\neq i} q_l^2 \texttt{Var}[B_{l,i}]$$
$$= q_i^2 \frac{2}{r} + \sum_{l\neq i} q_l^2 \frac{1}{r} = \frac{q_i^2}{r} + \sum_{l=1}^{m} q_l^2 \frac{1}{r} = \frac{q_i^2 + \sum_{l=1}^{m} q_l^2}{r} \approx \frac{\sum_{l=1}^{m} q_l^2}{r} = \frac{\|q\|_2^2}{r}, \tag{32}$$

where the approximation is indeed valid because m, which is the dimension of the input of the current layer (see Figure 1), is a large integer. Finally, according to Equation (31), each element u_i of qB is the sum of m random variables, for which the Lindeberg's condition is also satisfied: as $m \to \infty$, $s_m^2 = \frac{\|q\|_2^2}{r} \to \infty$ (m is the dimension of q, and s_m is the sum of variances of the m random variables, which we found in Equation (32)). Hence, $[|u_i - 0| > \epsilon s_m] \downarrow \emptyset$ as $m \to \infty$. Therefore, from Theorem D.4, we also conclude that as $m \to \infty$, each element of qB approaches a Gaussian with the mean and variance found in Equation (32). Therefore, we conclude that having an A, where the elements of A are *i.i.d* and from $\mathcal{N}(0, \frac{1}{r})$, then as $m \to \infty$, $q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$ approaches a Gaussian $\mathcal{N}(0, \frac{\|q\|^2}{r})$, which completes the proof.

□

F. Proof of Lemma 3.2

Despite having proved Lemma 3.1 when m approaches infinity, yet we need to quantify the distance between the distribution of $q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$ (for a given $q \in \mathbb{R}^{1\times m}$) and the Gaussian distribution with the same variance for limited values of m. In this section, we derive a closed form upper-bound for the total variation distance between the distribution of each element of $q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$ and the Gaussian distribution $\mathcal{N}(0, \|q\|^2/r)$.

Suppose $X_1, \dots, X_n$ are n independent random variables that are not necessarily *i.i.d*, and $\mathbb{E}[X_i] = 0$ (mean) and $\texttt{Var}[X_i] = \sigma_i^2 > 0$ (variance). Define $S_n = \sum_{i=1}^{n} X_i$ and let $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$. Assuming $Z_n = \frac{S_n}{s_n}$, and having Lindeberg's condition satisfied (see Theorem D.3 and Theorem D.4 in the appendix), the normalized sum Z_n has standard normal distribution in a weak sense for a bounded n. More precisely, the closeness of the cumulative distribution function (CDF) $F_n(x) = \texttt{Pr}\{Z_n \leq x\}$ to the standard normal CDF has been studied in terms of the Lyapounov ratios:

$$L_t = \frac{\sum_{i=1}^{n} \mathbb{E}[|X_i|^t]}{s_n^t}. \tag{33}$$

Particularly, if all $\{X_i\}_{i=1}^n$ are *i.i.d* and have a finite third absolute moment $\mathbb{E}[|X_i|^3]$, the classical Berry-Esseen theorem (gustav Esseen, 1945; Feller, 1971; Petrov, 1975) bounds the Kolmogrov distance between $F_n(x)$ and $\Phi(x)$:

$$\sup_x |F_n(x) - \Phi(x)| \leq CL_3, \tag{34}$$

where C is an absolute constant. In the more general case, when $\{X_i\}_{i=1}^n$ are independent (and not necessarily *i.i.d*), the number of summand variables n implicitly affects the value of L_3. Yet, we can bound the difference between $F_n(x)$ and $\Phi(x)$ in terms of generally stronger distances of total variation or entropic distance (Bobkov et al., 2011). To this end, let $D(X_i)$ denote the KL divergence between X_i and a normal variable from $\mathcal{N}(0, \sigma_i^2)$, i.e., the KL divergence between X_i and a normal variable with the same variance. Therefore, we have the following theorem about the distance between F_n and Φ:

Theorem F.1 ((Bobkov et al., 2011), theorem 1.1)**.** *Assume that the independent random variables $X_1, \ldots, X_n$ have finite third absolute moments, and that $D(X_i) \leq D$, where D is a non-negative number. Then,*

$$\|F_n(x) - \Phi(x)\|_{TV} \leq C_D L_3, \tag{35}$$

where the constant C_D depends on D only and $\|F_n(x) - \Phi(x)\|_{TV} = \sup_A \left| \int_A dF_n - \int_A d\Phi \right|$ is the total variation distance between F_n and Φ.

Having the theorem above, we can now derive a Berry-Esseen type bound for the total variation distance between each element of $q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$ in Lemma 3.1 and the normal law $\mathcal{N}(0, \|q\|^2/r)$: we need to find the third Lyapounov ratio for the summands contributing to each of the m elements in $q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$. To this end, we state and prove the following lemma:

Lemma F.2. *Let $A \in \mathbb{R}^{r\times m}$ be a matrix with i.i.d entries sampled from $\mathcal{N}(0, \frac{1}{r})$. Given a fixed $q \in \mathbb{R}^{1\times m}$ with $0 < c \leq |q_i| \leq C$, let $u = q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$. Let u_i be the i-th element of u and $Q_m(x) = \texttt{Pr}\{u_i \leq x\}$, the CDF of u_i. Also, let $\Phi(x)$ be the CDF of $z \sim \mathcal{N}(0, \frac{\|q\|^2}{r})$. Then:*

$$\|Q_m(x) - \Phi(x)\|_{TV} \in \mathcal{O}\left(\frac{1}{\sqrt{mr}}\right). \tag{36}$$

The above result shows that the distribution of $q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$ gets closer to $\mathcal{N}(0, \|q\|^2/r)$ as m and r increase. We now prove the above lemma before proceeding to the proof of Lemma 3.2.

Proof. From Equation (31), we had:

$$u_i = \sum_{l\neq i, l=1}^{m} q_l B_{l,i} + q_i B_{i,i}, \tag{37}$$

where $B_{l,i} = \frac{1}{r} A_{:,l}^\top A_{:,i} = \frac{1}{2r}\sum_{t=1}^r V_t$, where $V_t \sim \texttt{Variance-Gamma}(\nu, \alpha, \beta, \mu)$ with $\nu = \beta = \mu = 0$ and $\alpha = \frac{1}{2}$. Also $B_{i,i} = \frac{1}{r} A_{:,i}^\top A_{:,i} - 1 = \frac{\texttt{X}}{r} - 1$, where $X \sim \mathcal{X}_r^2$. Therefore, we can rewrite the equation above for u_i as:

$$u_i = \sum_{l\neq i, l=1}^{m} \frac{q_l}{2r} \sum_{t=1}^{r} V_t + q_i(\frac{\texttt{X}}{r} - 1) = \sum_{l\neq i, l=1}^{m} \sum_{t=1}^{r} \frac{q_l}{2r} V_t + \frac{q_i}{r}(\texttt{X} - r), \tag{38}$$

where $V_t \sim \texttt{Variance-Gamma}(\nu, \alpha, \beta, \mu)$ with $\nu = \beta = \mu = 0$ and $\alpha = \frac{1}{2}$ and $X \sim \mathcal{X}_r^2$. Hence, V_t has mean 0 and variance 4 and $(\texttt{X} - r)$ has mean 0 and variance $2r$. Also note that X can be written as the summation of r independent variables with distribution $\mathcal{X}_1^2$. Therefore, u_i is the weighted sum of mr independent random variables with mean 0. Also,

from Equation (32) in the proof of Lemma 3.1, we know that u_i has mean 0 and variance $\frac{\|q\|_2^2}{r}$. Now, in order to bound the TV distance between the distribution of u_i and $\mathcal{N}(0, \frac{\|q\|_2^2}{r})$, we have to use Theorem F.1 and Equation (33). More specifically, we have to find the third Lyapounov ratio $L_3 = \frac{\sum_i \mathbb{E}[|X_i|^3]}{s_n^3} = \frac{\sum_i \mathbb{E}[|X_i|^3]}{\left(\sum_i \texttt{Var}[X_i]\right)^3} = \frac{\sum_i \mathbb{E}[|X_i|^3]}{\left(\sum_i \mathbb{E}[X_i^2]\right)^3}$, where X_i is each of the $1+(m-1)r$ summands in Equation (38). First we note that, based on Equation (32), $s_n^3 = (\frac{\|q\|_2^2}{r})^{\frac{3}{2}} = \frac{\|q\|_2^3}{r\sqrt{r}}$. Now, we find the numerator $\sum_i \mathbb{E}[|X_i|^3]$. From (Gaunt, 2024), we know that for $V_t \sim \texttt{Variance-Gamma}(\nu, \alpha, 0, 0), \mathbb{E}[|V_t|^r] = \frac{2^r}{\sqrt{\pi}\alpha^r} \frac{\Gamma(\nu+(r+1)/2)\Gamma((r+1)/2)}{\Gamma(\nu+1/2)}$. Therefore, for $V_t \sim \texttt{Variance-Gamma}(0, \frac{1}{2}, 0, 0), \mathbb{E}[|V_t|^3] = \frac{2^6}{\pi}$. On the other hand, we know that the skewness of $X \sim \mathcal{X}_r^2$ is equal to $\frac{\mathbb{E}[(X-\mathbb{E}[X])^3]}{\texttt{Var}[X]^{\frac{3}{2}}} = \frac{\mathbb{E}[(X-r)^3]}{(2r)^{\frac{3}{2}}} = \sqrt{\frac{8}{r}}$. Hence, $\mathbb{E}[(X-r)^3] = (2r)^{\frac{3}{2}}\sqrt{\frac{8}{r}} = 8r$. Hence for $X \sim \mathcal{X}_r^2$, $\mathbb{E}[|X-r|^3] \geq \mathbb{E}[(X-r)^3] = 8r$. Now, we can find the numerator $\sum_i \mathbb{E}[|X_i|^3]$ as:

$$\begin{aligned}
\sum_i \mathbb{E}[|X_i|^3] &= \sum_{l\neq i, l=1}^{m} \sum_{t=1}^{r} \frac{|q_l|^3}{8r^3}\mathbb{E}[|V_t|^3] + \frac{|q_i|^3}{r^3}\mathbb{E}[|\texttt{X} - r|^3] \\
&= \sum_{l\neq i, l=1}^{m} \frac{|q_l|^3}{8r^2}\cdot\frac{2^6}{\pi} + \frac{|q_i|^3}{r^3}\mathbb{E}[|\texttt{X} - r|^3] \\
&\approx \sum_{l\neq i, l=1}^{m} \frac{8|q_l|^3}{\pi r^2} + \frac{8|q_i|^3}{r^2} \approx \sum_{l=1}^{m} \frac{8|q_l|^3}{\pi r^2} = \frac{8}{\pi r^2}\|q\|_3^3.
\end{aligned} \tag{39}$$

Therefore, for the sum u_i in Equation (38), we have the third Lyapounov ratio:

$$L_3 = \frac{8}{\pi r^2}\|q\|_3^3 \times \frac{r\sqrt{r}}{\|q\|_2^3} = \frac{8}{\pi\sqrt{r}}\left(\frac{\|q\|_3}{\|q\|_2}\right)^3. \tag{40}$$

Therefore, based on Theorem F.1, we have:

$$\|Q_m(x) - \Phi(x)\|_{TV} \leq \frac{8C_D}{\pi\sqrt{r}}\left(\frac{\|q\|_3}{\|q\|_2}\right)^3, \tag{41}$$

where $C_D \leq \frac{\pi\sqrt{r}}{8}$ is a constant, which depends only on D, where D is an upperbound for the KL divergence between each of the random variable summands in Equation (38) and a Gaussian with the same mean and variance. Now, assuming $0 < c \leq |q_i| \leq C$ for the elements q_i in q, we have $\left(\frac{\|q\|_3}{\|q\|_2}\right)^3 \leq \left(\frac{|C|}{|c|}\right)^3 \frac{1}{\sqrt{m}}$. Therefore:

$$\|Q_m(x) - \Phi(x)\|_{TV} \leq \frac{8C_D}{\pi}\left(\frac{|C|}{|c|}\right)^3 \frac{1}{\sqrt{mr}}. \tag{42}$$

Therefore,

$$\|Q_m(x) - \Phi(x)\|_{TV} \in \mathcal{O}\left(\frac{1}{\sqrt{mr}}\right). \tag{43}$$

□

We can now benefit from the useful coupling characterization of the total variation distance to establish a more understandable interpretation of the above result.

The coupling characterization of the total variation distance. For two distributions P and Q, a pair of random variables (X, Y), which are defined on the same probability space, is called a coupling for P and Q if $X \sim P$ and $Y \sim Q$ (Levin et al., 2008; Devroye et al., 2023). A very useful property of total variation distance is the coupling characterization (Levin et al., 2008):

$\|P - Q\|_{TV} \leq t$ if and only if there exists a coupling (X, Y) for them such that $\texttt{Pr}\{X \neq Y\} \leq t$.

We can now use the above coupling characterization to directly prove Lemma 3.2 from Lemma F.2 as follows:

Proof. Given a fixed $q \in \mathbb{R}^{1\times m}$ with $0 < c \leq |q_i| \leq C$, let $u = q \cdot (A^\top A - \mathbb{I}_m) \in \mathbb{R}^{1\times m}$. Let u_i be the i-th element of u. From Lemma F.2, we had the following upperbound for the total variation distance between $Q_m(x) = \texttt{Pr}\{u_i \leq x\}$ and $\Phi(x)$ (the CDF of $z \sim \mathcal{N}(0, \frac{\|q\|^2}{r})$):

$$\|Q_m(x) - \Phi(x)\|_{TV} \in \mathcal{O}\left(\frac{1}{\sqrt{mr}}\right). \tag{44}$$

Therefore, from the coupling characterization of the total variation distance we can directly conclude that there exists a coupling (u_i, z), where $z \sim \mathcal{N}(0, \frac{\|q\|^2}{r})$ and

$$\texttt{Pr}\{u_i \neq z\} \in \mathcal{O}\left(\frac{1}{\sqrt{mr}}\right). \tag{45}$$

□