# Poster: TensorUnlearn: Efficient Approximate Machine Unlearning with Kronecker-Factored Tensor Networks

Ali Mohammadi Ruzbahani
*Smart Cyber-Physical (SCPS) Lab*
*University of Calgary*
*Email: ali.mohammadiruzbaha@ucalgary.ca*

Abbas Yazdinejad
*DCAIlab*
*University of Regina*
*Email: Abbas.Yazdinejad@uregina.ca*

Hadis Karimipour
*Smart Cyber-Physical (SCPS) Lab*
*University of Calgary*
*Email: hadis.karimipour@ucalgary.ca*

***Abstract*—Machine unlearning is mandated by privacy regulations yet computationally challenging. Exact Newton-step unlearning requires Hessian inversion at $\mathcal{O}(n^3)$, infeasible for modern architectures. We propose TensorUnlearn, a two-stage framework that cascades Kronecker-Factored Approximate Curvature (KFAC) with Matrix Product Operator (MPO) tensor network decomposition to efficiently approximate the Fisher inverse. KFAC captures layer-wise curvature structure; MPO further compresses the resulting factors for wide layers where KFAC alone remains prohibitive. We derive explicit cascaded error bounds and provide $(\varepsilon, \delta)$-certified unlearning guarantees under a local quadratic regime, implying bounded membership-inference advantage in that regime.**

## 1. Introduction

The right to be forgotten (GDPR Article 17) requires removal of specific data influence from deployed models. Newton-step unlearning $\boldsymbol{\theta}_u = \boldsymbol{\theta}^* - H^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathcal{D}_f, \boldsymbol{\theta}^*)$ is principled but requires $\mathcal{O}(n^3)$ Hessian inversion, infeasible for $n > 10^6$ [1]. KFAC [2] decomposes each layer's Fisher as $F^{(\ell)} \approx A^{(\ell)} \otimes G^{(\ell)}$, reducing inversion to $\mathcal{O}(d_{\text{in}}^3 + d_{\text{out}}^3)$. However, for wide layers (e.g., VGG-11 on ImageNet-scale inputs where $d_{\text{in}} = 512{\times}7{\times}7 = 25088$: factor $A \in \mathbb{R}^{25088\times25088}$ requires $\sim 2.4\,\text{GB}$ storage and $\sim 1.6 \times 10^{13}$ FLOPs to invert), KFAC alone remains a bottleneck. Our CIFAR experiments, where VGG-11-BN operates at $d_{\text{in}} = 512$, serve as a controlled proof-of-concept; full validation at ImageNet scale is ongoing.

**Why tensor networks?** Matrix Product Operator (MPO) decomposition from quantum many-body physics compresses structured matrices by exploiting low-rank spectral decay and local correlations, properties we observe in KFAC factors of neural networks. By cascading KFAC with MPO, we reduce per-factor storage from $\mathcal{O}(d^2)$ to $\mathcal{O}(dr^2)$ and inversion cost to $\mathcal{O}(TKd_{\max}^2 r^3)$, making factor inversion sub-dominant to the forward pass.

## 2. Method and Theory

**Pipeline (Fig. 1).** Per layer $\ell$: (1) compute damped KFAC factors $\hat{A}^{(\ell)} = \frac{1}{N}\sum_i \mathbf{a}_i\mathbf{a}_i^\top + \lambda I$, $\hat{G}^{(\ell)} = \frac{1}{N}\sum_i \mathbf{g}_i\mathbf{g}_i^\top + \lambda I$; (2) for $d > 256$, compress each factor via TT-SVD with bond dimension $r$; (3) solve $\Delta W^{(\ell)} = (\tilde{G}^{(\ell)})^{-1}\nabla_{W^{(\ell)}}\mathcal{L}(\mathcal{D}_f, \boldsymbol{\theta}^*)(\tilde{A}^{(\ell)})^{-1}$ via ALS sweeping; apply $\boldsymbol{\theta}_u = \boldsymbol{\theta}^* - \Delta\boldsymbol{\theta} + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2 I)$.

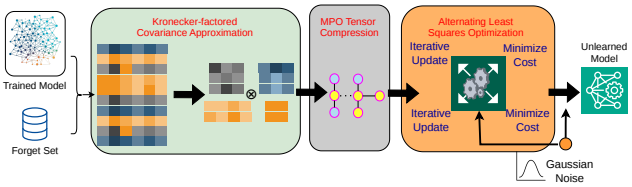



Figure 1. TensorUnlearn pipeline.

### 2.1. Cascaded Error Bound

**Theorem 1.** *Let $\tilde{F}^{(\ell)} = \tilde{A}^{(\ell)} \otimes \tilde{G}^{(\ell)}$ with MPO truncation errors $\epsilon_A^{(\ell)} = \left\|\hat{A}^{(\ell)} - \tilde{A}^{(\ell)}\right\|_F$, $\epsilon_G^{(\ell)} = \left\|\hat{G}^{(\ell)} - \tilde{G}^{(\ell)}\right\|_F$. The total approximation error decomposes as:*

$$\left\|F^{(\ell)} - \tilde{F}^{(\ell)}\right\|_F \le \underbrace{\epsilon_{KFAC}^{(\ell)}}_{\textit{Kronecker}} + \underbrace{\epsilon_A^{(\ell)}\left\|\hat{G}^{(\ell)}\right\|_F + (\left\|\hat{A}^{(\ell)}\right\|_F + \epsilon_A^{(\ell)})\epsilon_G^{(\ell)}}_{\textit{MPO truncation}}. \tag{1}$$

*Proof.* By triangle inequality: $\left\|F^{(\ell)} - \tilde{F}^{(\ell)}\right\|_F \le \left\|F^{(\ell)} - \hat{A}\otimes\hat{G}\right\|_F + \left\|\hat{A}\otimes\hat{G} - \tilde{A}\otimes\tilde{G}\right\|_F$. The first term is $\epsilon_{\text{KFAC}}^{(\ell)}$. For the second, write $\hat{A}\otimes\hat{G} - \tilde{A}\otimes\tilde{G} = (\hat{A}-\tilde{A})\otimes\hat{G} + \tilde{A}\otimes(\hat{G}-\tilde{G})$. By submultiplicativity of $\|\cdot\|_F$ under Kronecker products: $\le \epsilon_A\left\|\hat{G}\right\|_F + \left\|\tilde{A}\right\|_F\epsilon_G$. Substituting $\left\|\tilde{A}\right\|_F \le \left\|\hat{A}\right\|_F + \epsilon_A$ completes the proof. □

This bound gives *explicit knobs*: $\epsilon_{\text{KFAC}}$ is fixed by the network architecture; $\epsilon_A, \epsilon_G$ are controlled by bond dimension $r$, enabling principled accuracy-efficiency tradeoff.

### 2.2. Inverse Error Propagation

**Proposition 2.** *If $\hat{A}^{(\ell)}$ has condition number $\kappa_A$ and $\left\|\hat{A}-\tilde{A}\right\|_2 / \left\|\hat{A}\right\|_2 < 1/\kappa_A$, then:*

$$\left\|(\tilde{A}^{(\ell)})^{-1} - (\hat{A}^{(\ell)})^{-1}\right\|_2 \leq \frac{\kappa_A^2 \epsilon_A^{(\ell)}}{\left\|\hat{A}^{(\ell)}\right\|_2 (\kappa_A \left\|\hat{A}^{(\ell)}\right\|_2 - \kappa_A \epsilon_A^{(\ell)})}. \tag{2}$$

*An analogous bound holds for $\tilde{G}^{(\ell)}$. With damping $\lambda$, $\kappa_A \leq \left\|\hat{A}\right\|_2 / \lambda$, which is computable.*

*Proof.* By the Neumann series perturbation of matrix inverse: $\tilde{A}^{-1} - \hat{A}^{-1} = \hat{A}^{-1}(\hat{A} - \tilde{A})\tilde{A}^{-1}$. Taking norms and bounding $\left\|\tilde{A}^{-1}\right\|_2$ via the perturbation lemma for positive definite matrices yields the result. Damping ensures $\hat{A} \succeq \lambda I$, giving the computable bound on $\kappa_A$. □

### 2.3. Certified Unlearning (Local Regime)

Our certificates hold under a *local* $\mu$-strongly convex, $\beta$-smooth approximation of $\mathcal{L}$ near $\boldsymbol{\theta}^*$, as standard in influence-function-based analyses [3]; they should therefore be interpreted as *local certificates* rather than global guarantees for deep non-convex networks.

**Theorem 3.** *Let $\gamma = \max_\ell \left\|(\tilde{F}^{(\ell)})^{-1} - (F^{(\ell)})^{-1}\right\|_2$. TensorUnlearn with $\xi \sim \mathcal{N}(0, \sigma^2 I)$ achieves $(\varepsilon, \delta)$-certified unlearning [3] with:*

$$\varepsilon = \frac{\Delta_s^2}{2\sigma^2} + \frac{\Delta_s}{\sigma}\sqrt{2\ln\tfrac{1.25}{\delta}}, \tag{3}$$

*where $\Delta_s = \frac{|\mathcal{D}_f|}{\mu N} G_{\max} + \gamma \left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}_f, \boldsymbol{\theta}^*)\right\|_2$ and $G_{\max} = \max_{z \in \mathcal{D}_f} \left\|\nabla_{\boldsymbol{\theta}} \ell(z, \boldsymbol{\theta}^*)\right\|_2$.*

*Proof.* $\left\|\boldsymbol{\theta}_u - \boldsymbol{\theta}_r^*\right\|_2 \leq \underbrace{\gamma \left\|\nabla \mathcal{L}(\mathcal{D}_f)\right\|_2}_{\text{approx. gap}} + \underbrace{\frac{|\mathcal{D}_f|}{\mu N} G_{\max}}_{\text{Newton residual}} = \Delta_s$ by triangle inequality, where the Newton residual follows from strong convexity and Taylor remainder. The Gaussian mechanism with $\ell_2$-sensitivity $\Delta_s$ yields $(\varepsilon, \delta)$. □

**Corollary 4** (MIA Resilience)**.** *For any black-box MIA adversary $\mathcal{A}$ using loss-threshold or shadow-model attacks:* $\mathrm{Adv}(\mathcal{A}) \leq e^\varepsilon - 1 + \delta$.

**Complexity.** Per-layer cost is $\mathcal{O}(N(d_{\text{in}}^2 + d_{\text{out}}^2) + TK(d_{\max}^2 r^3 + d_{\max} r^4))$. With $T \leq 20$ ALS sweeps and $K \leq 4$, the second term is sub-dominant for $r \leq 64$.

## 3. Preliminary Results

Table 1 summarizes preliminary class-removal results. TensorUnlearn ($r$=32) achieves retain accuracy within 0.8% (CIFAR-10) and 0.8% (CIFAR-100) of full retraining at $\sim$15$\times$ and $\sim$14$\times$ speedup respectively. Certified $\varepsilon$=2.9 on CIFAR-10 compares favorably to Fisher Forgetting ($\varepsilon$=8.7).

TABLE 1. CLASS-REMOVAL UNLEARNING ACROSS DATASETS AND ARCHITECTURES. MIA DENOTES MIA SUCCESS RATE (IDEAL: 50%); $\varepsilon$ IS MEASURED AT $\delta = 10^{-5}$.

| Setting | Method | $\mathbf{Acc}_r \uparrow$ | $\mathbf{Acc}_f \downarrow$ | MIA | Time | $\varepsilon$ |
|---|---|---|---|---|---|---|
| **CIFAR-10** ResNet-18 $\lvert\mathcal{D}_f\rvert = 5$K | Retrain | 94.3 | 9.8 | 50.2 | 47m | 0 |
| | GA [4] | 89.1 | 6.2 | 54.8 | 35s | – |
| | FF [5] | 91.5 | 14.3 | 55.1 | 4.2m | 8.7 |
| | SCRUB [6] | 93.1 | 11.5 | 51.8 | 6.1m | – |
| | KFAC-only | 93.0 | 11.2 | 52.4 | 3.5m | 3.2 |
| | **Ours** ($r = 16$) | 93.2 | 11.0 | 52.1 | 2.4m | 3.1 |
| | **Ours** ($r = 32$) | 93.5 | 10.5 | 51.6 | 3.1m | 2.9 |
| **CIFAR-100** VGG-11-BN $\lvert\mathcal{D}_f\rvert = 500$ | Retrain | 71.8 | 0.8 | 50.1 | 38m | 0 |
| | GA | 65.2 | 0.3 | 56.4 | 28s | – |
| | FF | 68.4 | 3.1 | 55.8 | 3.8m | 11.2 |
| | SCRUB | 70.5 | 1.4 | 52.1 | 5.4m | – |
| | **Ours** ($r = 32$) | 71.0 | 1.1 | 51.4 | 2.8m | 3.7 |

The KFAC-only ablation isolates the MPO contribution: tensor compression adds $\sim$0.5% accuracy loss while providing $\sim$1.1$\times$ speedup on ResNet-18, with larger gains expected on wider architectures (VGG FC layers, ViT projections).

## 4. Conclusion

TensorUnlearn shows that cascading KFAC with MPO makes approximate Newton-style unlearning substantially more practical by compressing wide-layer curvature factors while retaining strong utility. Our results demonstrate consistent utility-privacy tradeoffs across CIFAR-scale settings and support the local-regime certified guarantees. Extending validation to ImageNet-scale and transformer architectures, and strengthening certification beyond the local quadratic regime, are important next steps.

## References


[1] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. ICML*, vol. 70, pp. 1885–1894, 2017.

[2] J. Martens and R. Grosse, "Optimizing neural networks with KFAC," in *Proc. ICML*, vol. 37, pp. 2408–2417, 2015. [3, 4]

[3] C. Guo, T. Goldstein, A. Hannun, and L. van der Maaten, "Certified data removal from machine learning models," in *Proc. ICML*, vol. 119, pp. 3832–3842, 2020.

[4] A. Thudi, G. Deza, V. Chandrasekaran, and N. Papernot, "Unrolling SGD: Understanding factors influencing machine unlearning," in *Proc. IEEE EuroS&P*, pp. 303–319, 2022. [5, 6]

[5] A. Golatkar, A. Achille, and S. Soatto, "Eternal sunshine of the spotless net: Selective forgetting in deep networks," in *Proc. CVPR*, pp. 9304–9312, 2020. [7, 8]

[6] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou, "Towards unbounded machine unlearning," in *Proc. NeurIPS*, 2023.

[7] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *Proc. IEEE CSF*, pp. 268–282, 2018.

[8] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *Proc. IEEE S&P*, pp. 1897–1914, 2022.

# TensorUnlearn: Efficient Approximate Machine Unlearning with Kronecker-Factored Tensor Networks

Ali Mohammadi Ruzbahani [1] Abbas Yazdinejad [2] Hadis Karimipour [1]

[1]Smart Cyber-Physical (SCPS) Lab, University of Calgary, {ali.mohammadiruzbaha, hadis.karimipour}@ucalgary.ca [2]DCAllab, University of Regina, Abbas.Yazdinejad@uregina.ca

## Motivation and Problem

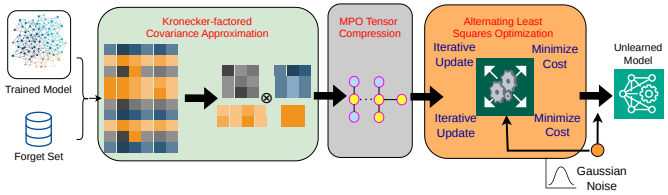
Privacy regulations such as GDPR Article 17 require model operators to remove specific training data influence upon request. The principled solution is Newton-step unlearning, $\boldsymbol{\theta}_u = \boldsymbol{\theta}^* - H^{-1}\nabla\mathcal{L}(\mathcal{D}_f, \boldsymbol{\theta}^*)$, but inverting the Hessian $H \in \mathbb{R}^{n\times n}$ is infeasible for modern architectures, ResNet-18 alone would need ∼512 TB just to store $H$. KFAC alleviates this by factoring each layer's Fisher as $A^{(\ell)} \otimes G^{(\ell)}$, yet for wide layers the factors themselves remain too large: VGG-11 on ImageNet-scale inputs ($d_{\text{in}} = 512{\times}7{\times}7 = 25{,}088$) produces a $25{,}088 \times 25{,}088$ factor requiring 2.4 GB and $\sim 10^{13}$ FLOPs to invert; our CIFAR experiments (where $d_{\text{in}} = 512$) serve as a controlled proof-of-concept. Existing approximate methods either lack formal privacy guarantees (gradient ascent, SCRUB) or cannot scale to these layer widths (full Fisher forgetting). No current approach simultaneously delivers tractable computation, near-retrain accuracy, and certified unlearning for architectures with wide layers.

**Research Question**

Can tensor network decomposition break the wide-layer bottleneck of curvature-based unlearning while preserving privacy certificates?

## TensorUnlearn: System Overview

Given a trained model and forget set $\mathcal{D}_f$, TensorUnlearn proceeds in three stages: (1) KFAC decomposes each layer's Fisher into Kronecker factors $A^{(\ell)} \otimes G^{(\ell)}$; (2) MPO compresses each factor into low-rank tensor cores with bond dimension $r$; (3) ALS sweeping solves the compressed inverse without reconstructing the full matrix. Calibrated Gaussian noise injected into the final update provides $(\varepsilon, \delta)$-certified guarantees.




## Methodology

We approximate the intractable Hessian inverse through two cascaded stages. First, KFAC decomposes each layer's Fisher into compact Kronecker factors from activation and gradient statistics:

$$F^{(\ell)} \approx \hat{A}^{(\ell)} \otimes \hat{G}^{(\ell)}, \qquad (F^{(\ell)})^{-1} \approx (\hat{A}^{(\ell)})^{-1} \otimes (\hat{G}^{(\ell)})^{-1}$$

For wide layers where these factors remain large, MPO tensor decomposition compresses each into a chain of low-rank cores with bond dimension $r$, reducing storage from $\mathcal{O}(d^2)$ to $\mathcal{O}(dr^2)$. The unlearning update is solved via ALS sweeping on the compressed factors with calibrated noise:

$$\boldsymbol{\theta}_u = \boldsymbol{\theta}^* - \bigoplus_{\ell} \text{vec}\Big((\tilde{G}^{(\ell)})^{-1} \nabla_{W^{(\ell)}} \mathcal{L}(\mathcal{D}_f, \boldsymbol{\theta}^*) \, (\tilde{A}^{(\ell)})^{-1}\Big) + \xi, \quad \xi \sim \mathcal{N}(0, \sigma^2 I)$$

The total approximation error decomposes into a Kronecker term (fixed by architecture) and an MPO term (controlled by $r$):

$$\|F^{(\ell)} - \tilde{A}^{(\ell)} \otimes \tilde{G}^{(\ell)}\|_F \;\le\; \underbrace{\epsilon^{(\ell)}_{\text{KFAC}}}_{\text{Stage 1}} + \underbrace{\epsilon^{(\ell)}_A \|\hat{G}^{(\ell)}\|_F + \big(\|\hat{A}^{(\ell)}\|_F + \epsilon^{(\ell)}_A\big)\, \epsilon^{(\ell)}_G}_{\text{Stage 2: MPO truncation}}$$

Under a local $\mu$-strongly convex, $\beta$-smooth approximation near $\boldsymbol{\theta}^*$ (standard in influence-function analyses; interpreted as local rather than global certificates for deep non-convex networks), TensorUnlearn achieves $(\varepsilon, \delta)$-certified unlearning with:

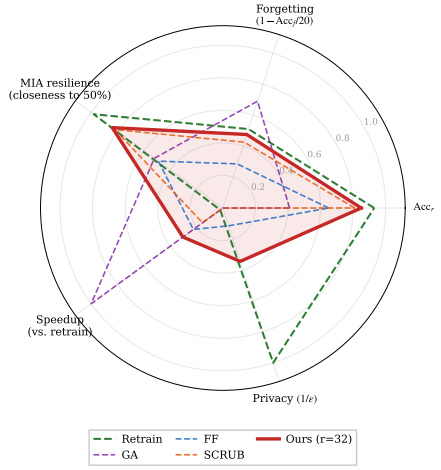
$$\varepsilon = \frac{\Delta_s^2}{2\sigma^2} + \frac{\Delta_s}{\sigma}\sqrt{2\ln\tfrac{1.25}{\delta}}, \qquad \Delta_s = \underbrace{\frac{|\mathcal{D}_f|}{\mu N} G_{\max}}_{\text{Newton residual}} + \underbrace{\gamma\, \|\nabla\mathcal{L}(\mathcal{D}_f, \boldsymbol{\theta}^*)\|_2}_{\text{KFAC-MPO gap}}$$

This directly bounds any black-box MIA adversary's advantage: $\text{Adv}(\mathcal{A}) \le e^{\varepsilon} - 1 + \delta$. Per-layer complexity is $\mathcal{O}\big(N(d_{\text{in}}^2 + d_{\text{out}}^2) + TK\, d_{\max}^2 r^3\big)$ with $T \le 20$ ALS sweeps.

## Experimental Setup

We evaluate on CIFAR-10 (ResNet-18, class removal, $|\mathcal{D}_f| = 5$K) and CIFAR-100 (VGG-11-BN, $|\mathcal{D}_f| = 500$) against Retrain (gold standard), Gradient Ascent, Fisher Forgetting, SCRUB, and a KFAC-only ablation. Metrics: retain accuracy ($\text{Acc}_r$), forget accuracy ($\text{Acc}_f \downarrow$), MIA success rate (ideal: 50%), wall-clock runtime, and certified $\varepsilon$ at $\delta = 10^{-5}$. Bond dimensions $r \in \{8, 16, 32\}$ are swept to characterize the accuracy-efficiency-privacy tradeoff; MIA uses loss-threshold and LiRA attacks under a black-box threat model.

## Preliminary Results




| Setting | Method | $\text{Acc}_r \uparrow$ | $\text{Acc}_f \downarrow$ | MIA | Time | $\varepsilon$ |
|---|---|---|---|---|---|---|
| CIFAR-10 ResNet-18 $\lvert\mathcal{D}_f\rvert = 5$K | Retrain | 94.3 | 9.8 | 50.2 | 47m | 0 |
| | GA | 89.1 | 6.2 | 54.8 | 35s | – |
| | FF | 91.5 | 14.3 | 55.1 | 4.2m | 8.7 |
| | SCRUB | 93.1 | 11.5 | 51.8 | 6.1m | – |
| | KFAC-only | 93.0 | 11.2 | 52.4 | 3.5m | 3.2 |
| | **Ours** ($r = 16$) | 93.2 | 11.0 | 52.1 | 2.4m | 3.1 |
| | **Ours** ($r = 32$) | 93.5 | 10.5 | 51.6 | 3.1m | 2.9 |
| CIFAR-100 VGG-11-BN $\lvert\mathcal{D}_f\rvert = 500$ | Retrain | 71.8 | 0.8 | 50.1 | 38m | 0 |
| | GA | 65.2 | 0.3 | 56.4 | 28s | – |
| | FF | 68.4 | 3.1 | 55.8 | 3.8m | 11.2 |
| | SCRUB | 70.5 | 1.4 | 52.1 | 5.4m | – |
| | **Ours** ($r = 32$) | 71.0 | 1.1 | 51.4 | 2.8m | 3.7 |

## Conclusion

TensorUnlearn demonstrates consistent utility-privacy tradeoffs across CIFAR-scale settings and supports local-regime certified guarantees. Extending validation to ImageNet-scale and transformer architectures, the wide-layer regime where MPO gains are theoretically largest, and strengthening certification beyond the local quadratic regime are important next steps.