GraSS: Scalable Data Attribution with Gradient Sparsification and Sparse Projection

Pingbang Hu[1] Joseph Melkonian[2] Weijing Tang[3] Han Zhao[1] Jiaqi W. Ma[1]

[1]University of Illinois Urbana-Champaign [2]Womp Labs [3]Carnegie Mellon University

Background: What is Data Attribution?

Given a dataset $D = \{z_i\}_{i=1}^n$ parametrized by a weight $w \in \mathbb{R}^n$, the corresponding model is trained via ERM $\mathcal{A}$ as:

$$\hat{\theta}_w = \mathcal{A}(w) := \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^n w_i \ell_i, \quad \ell_i := \ell(z_i; \theta).$$

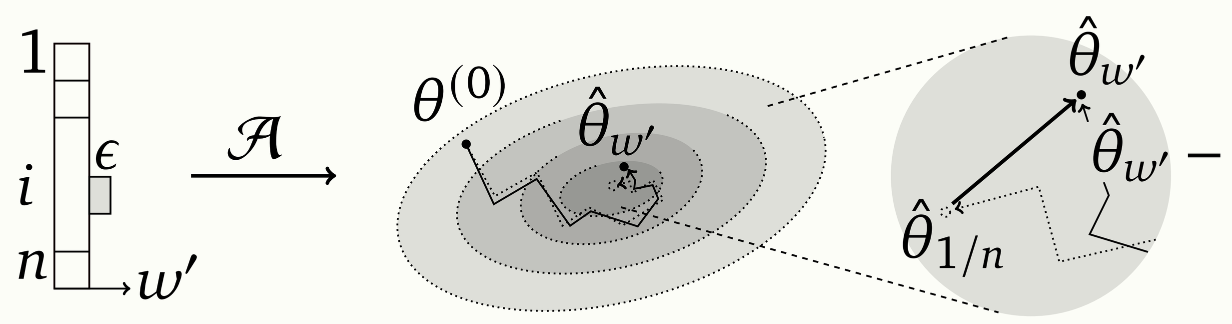
Default weight is $w = 1/n \in \mathbb{R}^p$, and we will first train $\hat{\theta}_{1/n}$.

Data attribution quantifies the **counterfactual effect** for *dataset perturbation* when w becomes w'. The key is to estimate $\hat{\theta}_{w'} - \hat{\theta}_w$.

Motivation: Gradient-Based Data Attribution

Most popular data attribution methods are gradient-based:

Intuition. Taylor-expand $\hat{\theta}_w$ around the default weight $1/n$ [2]:



$$\hat{\theta}_{w'} - \hat{\theta}_{1/n} \approx \epsilon \times -H^{-1}_{\hat{\theta}_{1/n}} \nabla_\theta \ell_i$$

Problem: Computing $H_{\hat{\theta}}^{-1} \nabla_\theta \ell_i$ is expensive, due to the size...

- Each $g_i := \nabla_\theta \ell_i$ is $\mathbb{R}^p$, and need inverting $H_{\hat{\theta}_{1/n}} \in \mathbb{R}^{p \times p}$.

Existing Approaches: Compression!

1. **Replace** $H_{\hat{\theta}}$ with *Fisher Information Matrix* $\frac{1}{n}\sum_{i=1}^n g_i g_i^\top \in \mathbb{R}^{p \times p}$.
2. **Compress** g_i from $\mathbb{R}^p$ to $\hat{g}_i \in \mathbb{R}^k$ with $k \ll p$!

$\Rightarrow$ FIM also reduces from $\mathbb{R}^{p \times p}$ down to $\mathbb{R}^{k \times k}$!

However, the **overhead of compression** is large:

- Dense matrix $P \in \mathbb{R}^{k \times p}$: $P g_i = \hat{g}_i$, $O(pk)$ per projection.
- SOTA (FJLT): $\widetilde{O}(p)$ per projection.
- SOTA (LoGra): $O(\sqrt{pk})$ per projection for **linear layers**.

Contributions

We design two *sub-linear* gradient compression algorithms:

1. GraSS: $O(k')$ per projection with $k < k' \ll p$.
2. FactGraSS: $O(k')$ but *without materializing* g_i for **linear layers**!

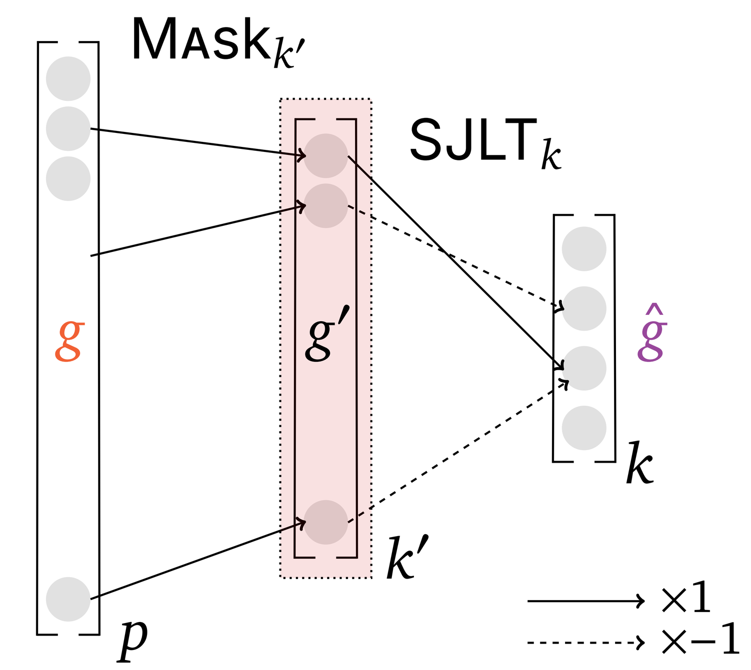
GraSS: Gradient Sparsification and Sparse Projection

GraSS compresses $g \in \mathbb{R}^p$ to $\hat{g} \in \mathbb{R}^k$ in $O(k')$ where $k < k' \ll p$:

Mask$_{k'}$. Sparsification:
- Select few parameters from g
- $\Rightarrow$ *Sub-linear* complexity!

+

SJLT$_k$. Sparse projection:
- Sparsify projection matrix P
- $\Rightarrow$ *Linear* complexity!



GraSS is already fast. But it requires materializing g.

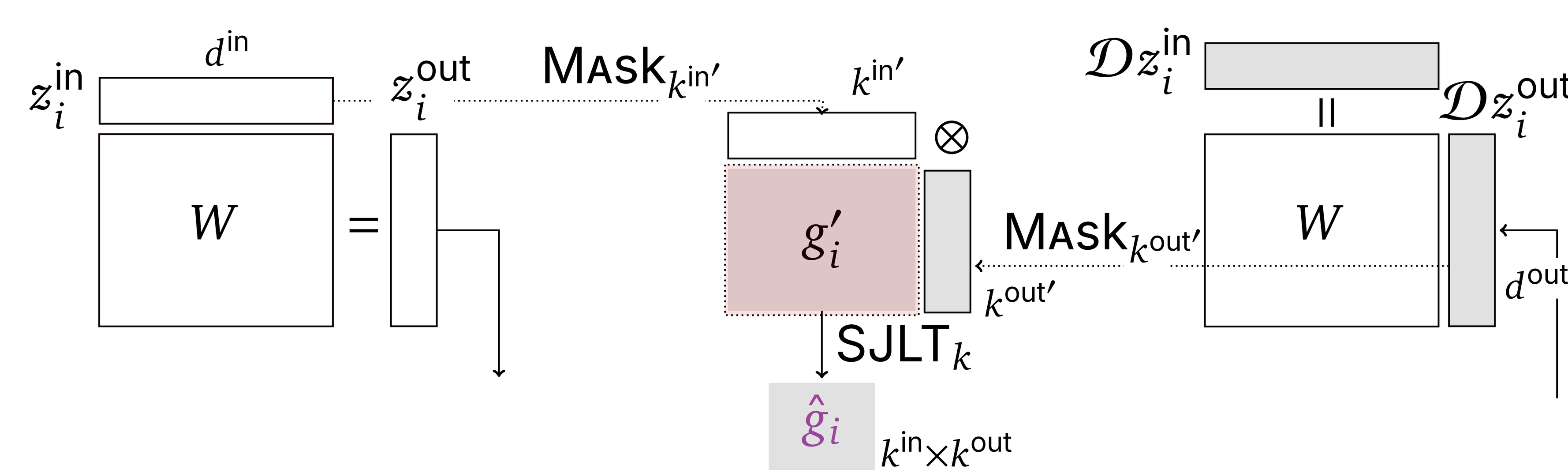
Q: Is this even a concern? **A:** Sadly, yes... Consider linear layers:

$$g_i = \frac{\partial \ell_i}{\partial W} = \frac{\partial \ell_i}{\partial z_i^{\text{out}}} \frac{\partial z_i^{\text{out}}}{\partial W} = z_i^{\text{in}} \otimes \frac{\partial \ell_i}{\partial z_i^{\text{out}}}$$

Previous SOTA gradient compression, LoGra [1], exploits this.

GraSS can also exploit this structure cleverly!

(1) **Factorized** Mask $\Rightarrow$ (2) Reconstruct $\Rightarrow$ (3) SJLT!



- **Bottlenecks**: SJLT's input size, $k' := k^{\text{in}\prime} \times k^{\text{out}\prime}$

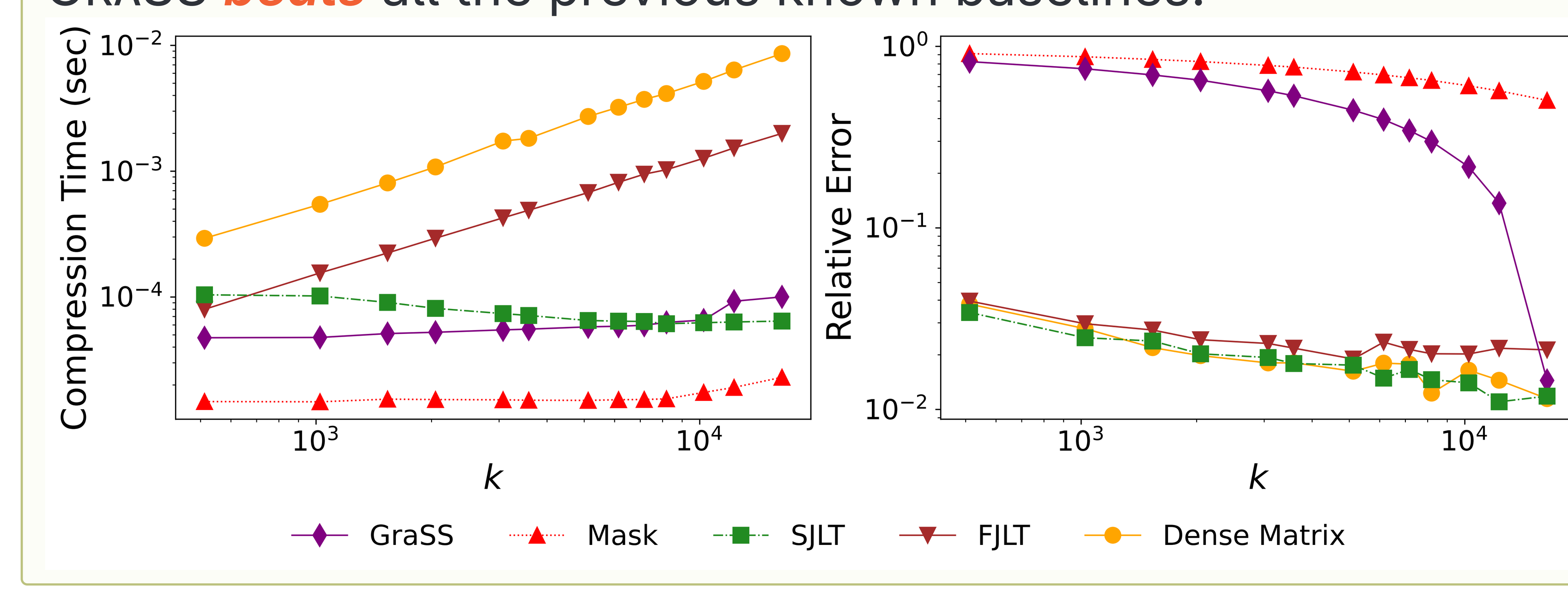
We summarize these two algorithms as follows:

Theorem. There is a *sub-linear* compression algorithm with complexity $O(k')$ where $k < k' \ll P$. Moreover, this extends to **linear layers**, where full gradients are **never** materialized.

Experimental Results

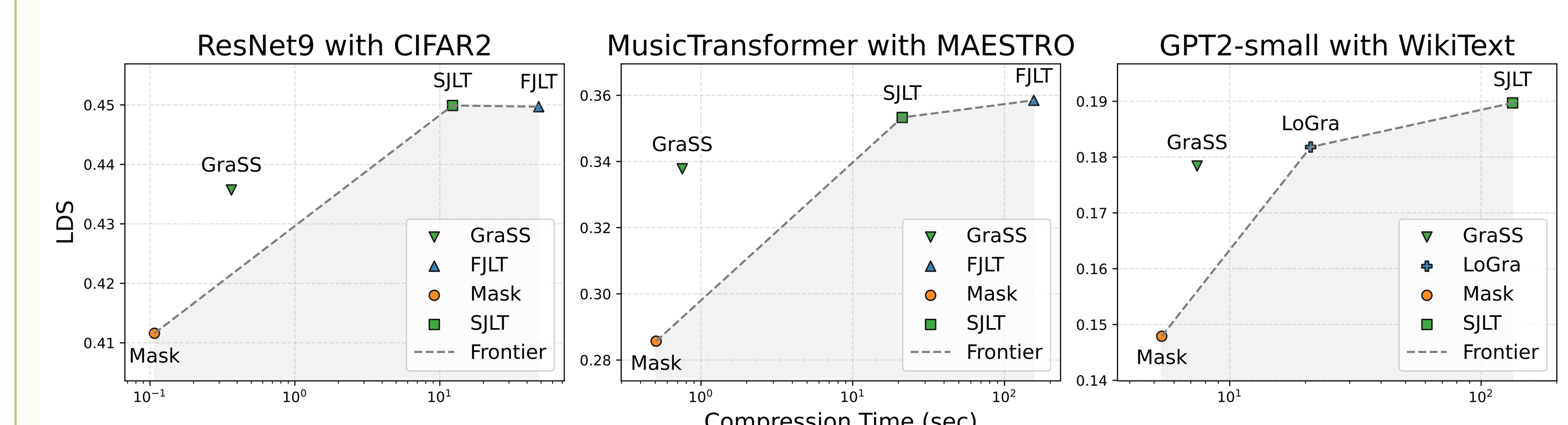
We first compare various baseline projectors on **general** inputs:

GraSS *beats* all the previous known baselines!



GraSS is fast, but not accurate. However, when on gradients:

New *Pareto frontier* on data attribution performance!



We scale the experiment to billion-scale model and datasets:

GraSS is *quantitatively accurate* on billion-scale!

To improve data privacy,

To improve data privacy, the European Union has implemented the General Data Protection Regulation (GDPR). ...
Data Protection Principles
The GDPR sets out six data protection principles...
- **Lawfulness, fairness, and transparency**: Businesses must process personal data in a way that is lawful, fair, and transparent. ...
- **Storage limitation**: Businesses must not store personal data for longer than necessary. ...

Data Subject Rights
The GDPR gives individuals a range of rights when it comes to their personal data. These rights include:
- **Right to access**: Individuals have the right to access their personal data and obtain information about how it is being processed. ...
- **Right to erasure**: Individuals have the right to have their personal data deleted if it is no longer necessary for the purposes for which it was collected. ...

Influential Data

...
The fact of registration and authorization of users on Sputnik websites via users' account or accounts on social networks indicates acceptance of these rules.
Users are obliged abide by national and international laws. ... The administration has the right to delete comments made in languages other than the language of the majority of the websites ...
...
- violates privacy, distributes personal data of third parties without their consent or violates privacy of correspondence; ...
- pursues commercial objectives, contains improper advertising unlawful political advertisement or links to other online resources ...

The administration has the right to block a user's access to the page or delete a user's account without notice if the user is in violation of these rules or if behavior indicating said violation is detected.
If the moderators deem it possible to restore the account/unlock access, it will be done. In the case of repeated violations of the rules above resulting in a second block of a user account, access cannot be restored. ...

[1] Choe et al. What is Your Data Worth to GPT? LLM-Scale Data Valuation with Influence Functions. 2025.

[2] Koh and Liang. Understanding black-box predictions via influence functions. PMLR. 2017.