Clustered Influence Functions

Amortized subset influence for high-query workflows

Miklós Máté Badó, Kristian Fenech

Eötvös Loránd University, Faculty of Informatics, Budapest, Hungary

Department of Artificial Intelligence | Neural Information Processing Group

bado@inf.elte.hu | fenech@inf.elte.hu

ICML 2026

Subset influence is expensive as a workload

The use case is not one query. It is many related subset queries.

Problem: Q queries $\rightarrow$ Q solves

$$S_q \mapsto H_\lambda^{-1} \sum_{i \in S_q} g_i, \qquad q = 1, \ldots, Q$$

$\Longrightarrow$

Observation

The curvature operator H_λ^{-1} is shared.

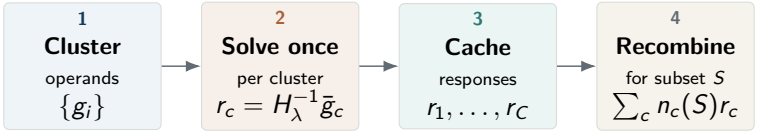
Only the operand changes:

$$v_q = \sum_{i \in S_q} g_i$$

CiF replaces repeated inverse-curvature solves with one reusable cache.

Cluster operands, cache inverse-curvature responses

At query time, CiF counts cluster memberships and recombines cached solves.



Approximation target

$$H_\lambda^{-1} \sum_{i \in S} g_i \;\approx\; \sum_{c=1}^{C} n_c(S)\, r_c$$

same damped-GGN IF target; operand space compressed

Q subset queries become C cached solves plus cheap recombination.

The approximation error has two sources

The bound separates modeling error from numerical error.

Cluster scatter

operand compression

Loose clusters give a poor surrogate for $\{g_i\}$.

Solver residual

IHVP accuracy

Loose solves make the responses inaccurate.

$$\|\text{CiF error}\| \lesssim \underbrace{\text{cluster scatter}}_{\text{increase } C} + \underbrace{\text{solver residual}}_{\text{tighten } \varepsilon}$$

The error bound tells which knob to turn.

CiF makes high-query subset influence feasible

CIFAR-10 fold workloads: one cache reused across 150 subset queries.

15.3×
LeNet
32.1 h → 2.1 h
$\rho_S = 0.94$

11.7×
AlexNet
63.3 h → 5.4 h
$\rho_S = 0.99$

13.9×
ResNet-18
75.0 h → 5.4 h
$\rho_S = 0.75$; 0.94 at $C = 50$

Q subset queries $\longrightarrow$ C cached solves + cheap recombination

Scalable influence analysis is a workload-level amortization problem.