Block-OBS-GS: Exact Per-Block Joint Brain Surgery with Gauss–Seidel Refinement for LLM Pruning

Yuwen Huang
The Hong Kong University of Science and Technology (Guangzhou)
yuwen.huang@ieee.org

Xiang Pan*
Lingnan University
xxiangpan@ln.edu.hk *Corresponding author

NeurIPS 2026 · Sydney

Scan for the paper (PDF)

Pruning can cut the storage and compute of deployed LLMs without retraining [1, 6].

Pruning an LLM block by block leaves the earlier blocks stale. One Gauss–Seidel sweep refreshes them.

- Converges to the global optimum with the mask fixed
- Same complexity class as SparseGPT
- 70B: SparseGPT's time, lower perplexity

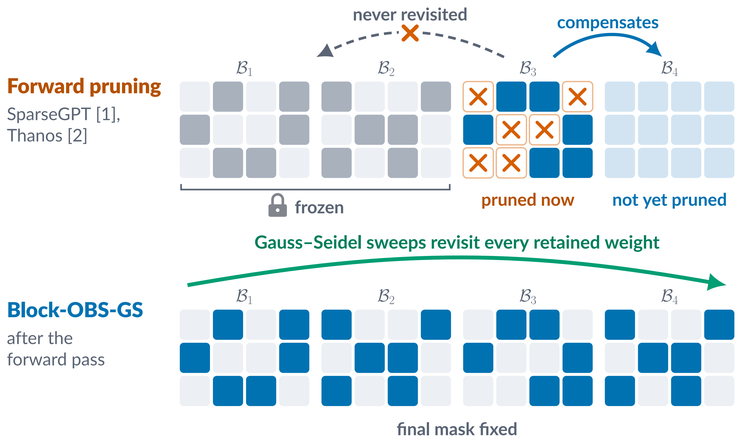
−14% — Llama-3.1-8B, 75%
perplexity vs SparseGPT
36.86 → 31.69

−47% — Llama-3.1-8B, 87.5%
perplexity vs SparseGPT
256.7 → 137.2

12/12 — same-mask settings
one GS sweep beats SparseGPT reconstruction

70B — same time as SparseGPT
perplexity 6.039 vs 6.080
time 4,313 vs 4,248 s

1 The problem: frozen blocks



For a layer $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ with calibration inputs $\boldsymbol{X} \in \mathbb{R}^{n\times T}$, each row i keeps its pruned weights at zero and minimizes

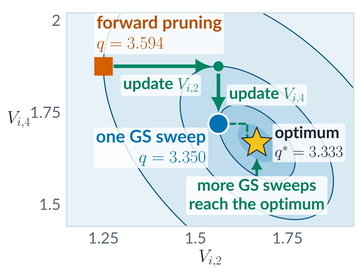
$$q(\boldsymbol{V}_{i,:}) \triangleq (\boldsymbol{V}_{i,:} - \boldsymbol{W}_{i,:}) \cdot \boldsymbol{H} \cdot (\boldsymbol{V}_{i,:} - \boldsymbol{W}_{i,:})^\top, \qquad \boldsymbol{H} \triangleq c \cdot \boldsymbol{X} \cdot \boldsymbol{X}^\top + \lambda \cdot \boldsymbol{I}.$$

Factor-graph view [5]: the forward pass decides too early



(a) forward pass fixes block $\mathcal{B}$ (b) GS sweeps on the complete graph

Circles: weights; squares: factors of $p \propto e^{-q}$, with the zero constraints δ_j in orange.



A 4-weight example

$\boldsymbol{W}_{i,:} = (1, 1, 1, 1)$, $\boldsymbol{H} = \boldsymbol{I} + \boldsymbol{1}\cdot\boldsymbol{1}^\top$; weights 1 and 3 deleted; blocks $\{1, 2\}$ and $\{3, 4\}$.

15/16

of the gap to the optimum is closed by one GS sweep

Forward pruning leaves $\partial q/\partial V_{i,2} = -5/4$ at the frozen weight $V_{i,2}$.

2 Our fix: solve, then sweep

STAGE 1 — **Forward block solve**
block by block, one exact solve per row

→

STAGE 2 — **Gauss–Seidel sweeps**
mask fixed, every block revisited

Stage 1: exact joint solve per row and block (multi-weight OBS [3])

$$\boldsymbol{V}^{(r)}_{i,\mathcal{B}\cup\mathcal{R}} = \boldsymbol{V}^{(r-1)}_{i,\mathcal{B}\cup\mathcal{R}} - \boldsymbol{V}^{(r-1)}_{i,\mathcal{D}_\mathcal{B}} \cdot \left((\boldsymbol{F}_{\mathcal{B},\mathcal{D}_\mathcal{B}})^\top \cdot \boldsymbol{F}_{\mathcal{B},\mathcal{D}_\mathcal{B}}\right)^{-1} \cdot (\boldsymbol{F}_{\mathcal{B},\mathcal{D}_\mathcal{B}})^\top \cdot \boldsymbol{F}_{\mathcal{B},\mathcal{B}\cup\mathcal{R}}$$

$\boldsymbol{F}$: upper Cholesky factor of $\boldsymbol{H}^{-1}$; $\mathcal{D}_\mathcal{B}$: deleted columns of block $\mathcal{B}$; $\mathcal{R}$: later columns.

Stage 2: Gauss–Seidel update of each retained weight, $j = 1, \ldots, n$

$$V^{(t)}_{i,j} \triangleq V^{(t-1)}_{i,j} - \frac{1}{H_{j,j}} \cdot \underbrace{\left(\sum_{l<j} H_{j,l} \cdot \left(V^{(t)}_{i,l} - W_{i,l}\right) + \sum_{l\geq j} H_{j,l} \cdot \left(V^{(t-1)}_{i,l} - W_{i,l}\right)\right)}_{\triangleq R_{j,i}}$$

Deleted weights stay zero; $R_{j,i}$ is the (j, i) entry of $\boldsymbol{R} = \boldsymbol{H} \cdot (\boldsymbol{V} - \boldsymbol{W})^\top$, kept current by one rank-one update per column.

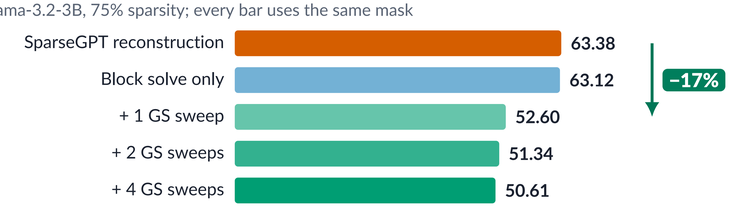
Guarantees

- **Exact block solve.** The Stage 1 update is the unique minimizer of q over $\boldsymbol{V}_{i,\mathcal{B}\cup\mathcal{R}}$ with $\boldsymbol{V}_{i,\mathcal{D}_\mathcal{B}} = 0$ and earlier blocks fixed.
- **Converges to the global optimum with the mask fixed.** $\|\boldsymbol{e}^{(K)}\|_{\boldsymbol{H}_{\mathcal{U}\mathcal{U}}} \leq \gamma^K \cdot \|\boldsymbol{e}^{(0)}\|_{\boldsymbol{H}_{\mathcal{U}\mathcal{U}}}$ with $\gamma < 1$: the sweeps converge geometrically to the global optimum $\boldsymbol{V}^*$.
- **Built-in certificate.** $q(\boldsymbol{V}_{i,:}) - q(\boldsymbol{V}^*_{i,:}) \leq \|\boldsymbol{R}_{\mathcal{U},i}\|_2^2/\lambda$.
- **MAP view.** Minimizing q is maximum a posteriori (MAP) estimation for the log-concave $p \propto e^{-q}$; each GS update is an iterated-conditional-modes step [4] on its factor graph [5].
- **Same complexity class as SparseGPT.** $\Theta(n^3 + n^2 \cdot T + m \cdot n^2)$ per layer; measured time 1.02× SparseGPT's at 70B and 2.0× at 8B.

$\mathcal{U}$: retained columns of row i; V^*: global optimum with the mask fixed; $e^{(K)}$: error of $\boldsymbol{V}_{i,\mathcal{U}}$ after K sweeps.

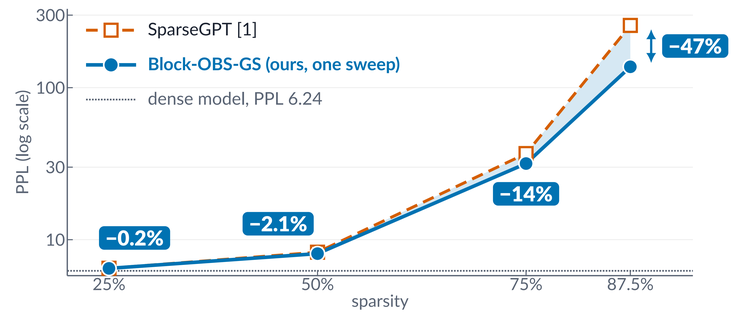
GS does the work

Llama-3.2-3B, 75% sparsity; every bar uses the same mask



3 Results

Llama-3.1-8B: perplexity (PPL) on WikiText-2, lower is better



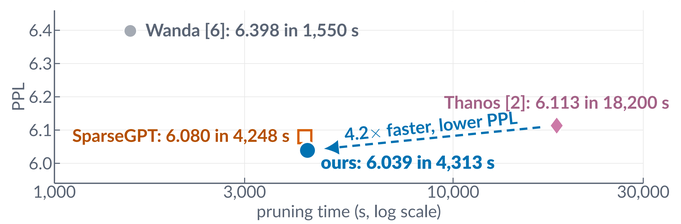
One GS sweep wins in all 12 settings

PPL change vs SparseGPT reconstruction on the same mask; darker means a larger gain

	25%	37.5%	50%	62.5%	75%	87.5%
Llama-3.2-3B	−0.3%	−0.9%	−3.1%	−9.1%	−17.0%	−10.8%
Llama-3.1-8B	−0.2%	−0.7%	−1.9%	−6.2%	−12.7%	−21.1%

70B scale: best PPL at SparseGPT's speed

Llama-3.3-70B-Instruct, 50%, six GPUs; dense PPL 3.862; block width 128 (ours), 32 (SparseGPT, Thanos)



Drop-in for compression pipelines

PPL, SparseGPT → Block-OBS-GS

−12.7% — on a fixed Wanda mask [6], 3B
75%: 182.95 → 159.63

−10.5% — on a fixed Wanda mask [6], 8B
75%: 124.10 → 111.04

−4.7% — followed by 4-bit GPTQ [7], 3B
50%: 16.13 → 15.37

References. [1] E. Frantar, D. Alistarh. SparseGPT. ICML 2023. [2] I. Ilin, P. Richtárik. Thanos. arXiv:2504.05346, 2025. [3] B. Hassibi, D. G. Stork. Second order derivatives for network pruning: Optimal Brain Surgeon. NeurIPS 5, 1992. [4] J. Besag. On the statistical analysis of dirty pictures. J. R. Stat. Soc. B, 1986. [5] F. R. Kschischang, B. J. Frey, H.-A. Loeliger. Factor graphs and the sum-product algorithm. IEEE Trans. Inf. Theory, 2001. [6] M. Sun, Z. Liu, A. Bair, J. Z. Kolter. A simple and effective pruning approach for large language models. ICLR 2024. [7] E. Frantar, S. Ashkboos, T. Hoefler, D. Alistarh. GPTQ. ICLR 2023.

Acknowledgments. Supported in part by the National Natural Science Foundation of China (62601349), the Research Grants Council of Hong Kong (23201626), HKUST(GZ) (G0101000351), and Lingnan University (SDS24A16).