The Iterative Optimal Brain Surgeon (I-OBS)

Faster Sparse Recovery by Leveraging Second-Order Information

Diyuan Wu[1] **Ionut-Vlad Modoranu**[1] **Mher Safaryan**[1]
Denis Kuznedelev[2,3] **Dan Alistarh**[1]

[1]Institute of Science and Technology Austria (ISTA) [2]Yandex Research [3]Skoltech

November 11, 2024

Motivation - sparse optimization problem

Sparse optimization problem

$$\min_{\theta \in \mathbb{R}^d} f(\theta) \quad \text{subject to} \quad \|\theta\|_0 \leq k$$

First-order methods: k-IHT

$$\theta_{t+1} = T_k(\theta_t - \eta \nabla f(\theta_t)), \tag{1}$$

where the operator T_k denote the top-k operator.

k-IHT as proximal point method

(1) can be considered as the solution of:

$$\theta_{t+1} = \arg\min_{\theta : \|\theta\|_0 \leq k} \langle \nabla f(\theta_t), \theta - \theta_t \rangle + \frac{1}{2}\|\theta - \theta_t\|_2^2.$$

Second-order method

Second-order method as proximal point method

$$\theta_{t+1} = \underset{\theta:\|\theta\|_0 \leq k}{\arg\min} \ \langle \nabla f(\theta_t), \theta - \theta_t \rangle + \frac{1}{2}\|\theta - \theta_t\|^2_{\boldsymbol{H}_t}. \tag{2}$$

where $\boldsymbol{H}_t = \nabla^2 f(\theta_t)$

Key contributions

- Derive I-OBS algorithm by solving (2).
- Prove local convergence rates under strongly convex and smooth assumptions.
- Applied the methods to model pruning problem.

The I-OBS algorithm - theoretical optimal version

Theoretical optimal iteration

- Compute dense Newton's update $\theta_t^+ = \theta_t - \mathbf{H}_t^{-1}\nabla f(\theta_t)$.
- Solve the optimal mask Q_{t+1}: with $\mathbf{H}_t^S := \mathbf{I}_S^\top \left(\mathbf{I}_S \mathbf{H}_t^{-1} \mathbf{I}_S^\top\right)^{-1} \mathbf{I}_S$, set $Q_{t+1} = [d] \setminus S_{t+1}$ where

$$S_{t+1} = \operatorname*{arg\,min}_{S:|S|=d-k} (\theta_t^+)^\top \mathbf{H}_t^S (\theta_t^+),$$

- Update the parameters: $\theta_{t+1} = \left(\mathbf{I} - \mathbf{H}_t^{-1}\mathbf{H}_t^{S_{t+1}}\right)\theta_t^+$

Practical iteration

- Compute dense Newton's update $\theta_t^+ = \theta_t - \mathbf{H}_t^{-1}\nabla f(\theta_t)$.
- Use the top-k mask Q_{t+1}: with $Q_{t+1} = \operatorname{supp} T_k(\theta_t^+)$
- Update the parameters: $\theta_{t+1} = (\theta_t^+)_{Q_{t+1}} = T_k(\theta_t^+)$

The I-OBS algorithm - rate of convergence

Rate of convergence

Assume θ_* is the unique k_*-sparse solution, and strong convexity, first-, second-order smoothness of f, both the theoretical and practical iteration satisfies:

$$\|\theta_{t+1} - \theta_*\|_2 \leq C\|\theta_t - \theta_*\|_2^2$$

Local quadratic convergence

The above rate of convergence implies $\mathcal{O}(\log\log\frac{1}{\epsilon})$ iteration complexity once $\|\theta_t - \theta_*\|_2 \leq \frac{1}{2c}$

Application to model pruning

Pruning methods as special case of I-OBS

WoodFisher/WoodTaylor [1] and OBC [2] are special case of I-OBS

[1] Sidak Pal Singh, and Dan Alistarh. "Woodfisher: Efficient second-order approximation for neural network compression."

[2] Elias Frantar , and Dan Alistarh. "Optimal brain compression: A framework for accurate post-training quantization and pruning."

Application to model pruning

I-OBS as iterative pruning method

Input: Sparsity threshold $k_\ell \in [d]$ for each layer $\ell \in \{1, 2, \ldots, L\}$
for each round $t \in \{1, 2, \ldots, T\}$ **do**
    Sample a data batch X_t, $H^0 \leftarrow X_t$
    for each layer $\ell \in \{1, 2, \ldots, L\}$ **do**
        Solve the constrained optimization problem

$$\min_{\widehat{W}^\ell} \| W_{t-1}^\ell H^{\ell-1} - \widehat{W}^\ell H^{\ell-1} \|_2^2 \quad s.t. \ \|\widehat{W}^\ell\|_0 = k_\ell$$

        using OBC or SparseGPT.
        $W_{t-1}^\ell \leftarrow \widehat{W}^\ell$
        $W_t^\ell \leftarrow W_{t-1}^\ell - \eta g_t^\ell(X_t, W_{t-1}^\ell)$ for each $\ell \in \{1, 2, \ldots, L\}$
    end for
end for

Experimental results for model pruning

Table: Pruning results for Phi-1.5M using SparseGPT. We report perplexity (the lower, the better).

Model	# samples	WikiText2			C4		
		Dense	SparseGPT	I-OBS(3)	Dense	SparseGPT	I-OBS(3)
OPT-125M	128	27.65	33.85	25.20	24.61	32.27	31.41
Phi-1.5	128	21.82	25.28	23.94	20.90	21.13	20.26

Experimental results on model pruning

Table: Performance of I-OBS on Llama-2-7b

Iterations	**MMLU(5-shot)**
0 (dense)	0.4584
1	0.3878
2	**0.3950**
3	0.3932
4	0.3943
5	0.3946
6	0.3929
7	0.3919
8	0.3893
9	0.3903
10	0.3863