

### LIMITED-MEMORY STOCHASTIC PARTITIONED QUASI-NEWTON TRAINING

Paul Raynaud, Dominique Orban

#### ▶ To cite this version:

Paul Raynaud, Dominique Orban. LIMITED-MEMORY STOCHASTIC PARTITIONED QUASI-NEWTON TRAINING. Edge Intelligence Workshop 2022, Sep 2022, Montréal Québec, Canada. hal-04062882

#### HAL Id: hal-04062882 https://hal.science/hal-04062882

Submitted on 7 Apr 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

# **GERAD** LIMITED-MEMORY STOCHASTIC PARTITIONED QUASI-NEWTON TRAINING



{ PAUL RAYNAUD, DOMINIQUE ORBAN } GERAD, POLYTECHNIQUE MONTRÉAL

## PARTIALLY-SEPARABLE TRAINING PROBLEM

Loss structure

$$\min_{w \in \mathbb{R}^n} \frac{1}{N} \sum_{l=1}^N \mathcal{L}(x^{(l)}, y^{(l)}; w),$$

resembles that of partially-separable  $f : \mathbb{R}^n \to \mathbb{R}$ 



where the  $U_j \in \mathbb{R}^{n_j \times n}$  select a subset of variables. The negative log likelihood

based on class scores  $c_j$ ,  $1 \le j \le C$ , is not partially-separable. We define the partially-separable loss (PSL)





LENET ARCHITECTURE

# PARTITIONED ARCHITECTURE

- Split two consecutive layers into *C* densely-connected partitions.
- Involve *C* times fewer variables than dense layers.
- Each score depends on a small fraction of variables.



 $h_{k,j}(w) := \frac{1}{N} \sum_{l=1}^{N} \delta_{y^{(l)},k} e^{c_j(x^{(l)};w) - c_k(x^{(l)};w)},$ 

where  $\delta_{y^{(l)},k} = 1$  if  $y^{(l)} = k$ , and 0 otherwise. Each element depends of one pair of scores.

# PARTITIONED QUASI-NEWTON METHODS

The partitioned structure of derivatives of Papartially-separable *f* allow us to constructs  $B \approx \nabla^2 f(w)$  as

 $B := \sum_{j=1}^{N} U_j^{\top} B_j U_j,$ 

where  $B_j \approx \nabla^2 f_j(w)$  is a quasi-Newton linear operator. Example:

- f Partitioned approximations:
  - are finer Hessian approximations;
  - are limited memory operators: storage in  $O(\sum_{j=1}^{N} 2mn_j), 1 \le m \le 5;$

• produce 3 methods:

– PLBFGS: each  $B_j$  is a LBFGS operator;



• LeNet's scores are parametrized by every variable beneath the last layer.

### **ARCHITECTURE DETAILS**

PSNet (n=12200)LeNet (n=24092)Conv 10 channelsConv 6 channels4x4 kernel, stride=1, max-pooling 2x2Conv 20 channelsConv 6 channels4x4 kernel, stride=1, max-pooling 2x2Separable 320x200Dense 256x84Separable 200x100Dense 84x10Separable 100x10



Figure 1: PSNet architecture

- Must not be terminated by a dense layer.
- As the number of scores increases, the fraction of variables they depend on decreases.



- PLSR1: each  $B_j$  is a LSR1 operator;
- PLSE: mixes LBFGS and LSR1 operators to best satisfy the secant equation  $B_{k+1}(w_{k+1} - w_k) = \nabla f(w_{k+1}) - \nabla f(w_k)$ .

# FUTURE RESEARCH

- Improve partially-separable networks evaluations, in particular the partitioned gradient computation or adapt the partitioned updates to use an aggregate gradient for the variables shared.
- Extend the partial separability concepts to other networks (mainly residual neural network).
- Develop parallel partitioned methods to run on several GPUs simultaneously.
- Explore how layer's dropout during training reduce the overlapping between element functions.
- Study how separable layers affect the vanishing gradient problem.

### RESULTS

# PSNet vs. LeNet on MNIST



# **CONTACT INFORMATION**

Web github.com/paraynaud,
 dpo.github.io
Email paul.raynaud@polymtl.ca,
 dominique.orban@gerad.ca

Code and more

github.com/JuliaSmoothOptimizers



Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. <u>Proceedings of the IEEE</u>, 86(11):2278–2324, Nov 1998.

[2] A Griewank and Ph. L. Toint. Partitioned variable metric updates for large structured optimization problems. <u>Numer. Math.</u>, 39:119–137, 1982. • Close accuracy between LeNet and PSNet.

• Similar asymptotic accuracy for all methods.

• Noisy progress of partitioned quasi-Newton method accuracies.