LIMITED-MEMORY STOCHASTIC PARTITIONED QUASI-NEWTON TRAINING

Paul Raynaud, Dominique Orban

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 $f : \mathbb{R}^n \to \mathbb{R}$:

$$f(w) = \sum_{j=1}^{N} f_j(U_j w), \quad f_j : \mathbb{R}^n_j \to \mathbb{R}, \quad n_j < n,$$

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

$$-\log\left(e^{c_{y^{(l)}}(x^{(l)};w)} / \sum_{j=1}^{C} e^{c_j(x^{(l)};w)}\right),$$

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

$$\begin{aligned}\mathcal{L}(w) &:= \frac{1}{N}\sum_{l=1}^{N}\sum_{j=1}^{C} e^{c_j(x^{(l)};w) - c_{y^{(l)}}(x^{(l)};w)},\\ &= \sum_{k=1}^{C}\sum_{j=1\neq k}^{C} h_{k,j}(w)\end{aligned}$$

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

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

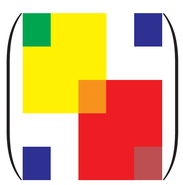
Partitioned quasi-Newton methods

The partitioned structure of derivatives of partially-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:

$$f(w) = f_1(w_1, w_2, w_3) + f_2(w_3, w_4, w_5) + f_3(w_1, w_5)$$

$B =$ 

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;
  - 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.

Contact Information

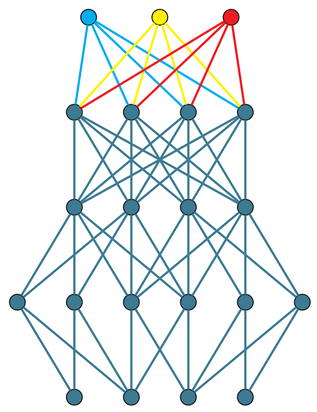
Web github.com/paraynaud, dpo.github.io
Email paul.raynaud@polymtl.ca, dominique.orban@gerad.ca
Code and more github.com/JuliaSmoothOptimizers

References

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

[2] A Griewank and Ph. L. Toint. Partitioned variable metric updates for large structured optimization problems. Numer. Math., 39:119–137, 1982.

LeNet architecture



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

Architecture details

PSNet (n=12200)	LeNet (n=24092)
Conv 10 channels	Conv 6 channels
4x4 kernel, stride=1, max-pooling 2x2	
Conv 20 channels	Conv 6 channels
4x4 kernel, stride=1, max-pooling 2x2	
Separable 320x200	Dense 256x84
Separable 200x100	Dense 84x10
Separable 100x10	

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.

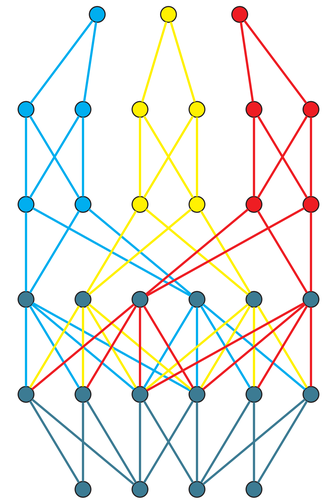
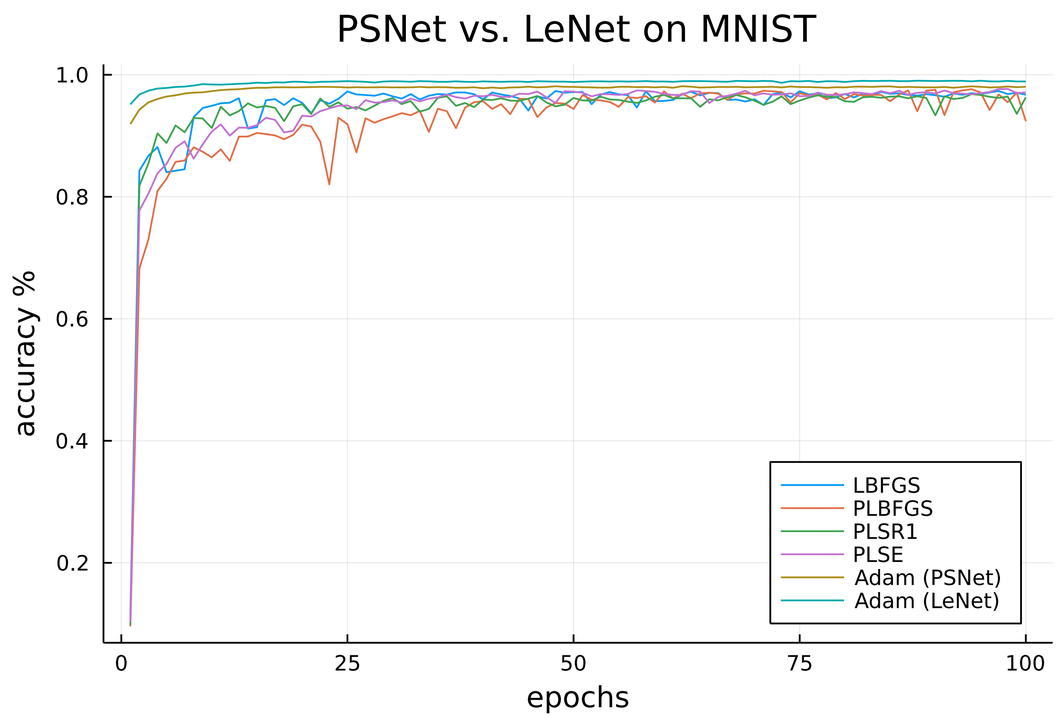


Figure 1: PSNet architecture

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

Results



- Close accuracy between LeNet and PSNet.
- Similar asymptotic accuracy for all methods.
- Noisy progress of partitioned quasi-Newton method accuracies.