ENSAE ParisTech
École nationale
de la statistique
et de l'administration
économique
université
PARIS-SACLAY

Rapport SOADML : ***Sujet 2***

NON CONVEX FINITE SUM OPTIMIZATION VIA SCSG METHODS

Auteurs :
Kossi NEROMA
Marin Boutemy

Chargés du cours :
Marco Cuturi

23 Mai 2019

Contents

1 The SCSG algorithm

The stochastically controlled stochastic gradient (SCSG) methods belong to a class of algorithm introduced by [L. Lei, 2016] and designed to tackle smooth non-convex finite sum optimization problems. They are proved to outperform the classical stochastic gradient descent (SGD) algorithm if the smoothness of each component of the finite sum holds.

The general form of a finite sum optimization problem is as follows :

$$\min_{w \in \mathbb{R}^d} f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) \tag{1}$$

The proposed algorithm's pseudo-code is a kind of a mix between `mini-batch SGD`[Bottou, 2010] algorithm and the SVRG [L. Lei, 2016] one. But, instead of making a random pass over the data at each step, SCSG draws a geometrically controled random number of mini-batches. According to our base paper [L. Lei, 2017] results, this scheme would yield to a faster convergence and a more robust estimator.

Algorithm 1 SCSG algorithm

T

```
procedure SCSGMethod($T$, $\tilde{w}_0$, $\mathbf{B}$, $\mathbf{b}$, $\eta$)
    $\mathbf{w} \leftarrow w_0$
    for $j \leftarrow 1, T$ do
        Uniformly sample a batch $\mathcal{I}_j \subset \{1, ..., n\}$ with $\mid \mathcal{I}_j \mid = B_j$
        $g_j \leftarrow \nabla f_{\mathcal{I}_j}(\tilde{w}_{j-1})$
        $w_0^{(j)} \leftarrow \tilde{w}_{j-1}$
        Generate $N_j \sim \text{Geo}(B_j/(B_j + b_j))$
        for $k \leftarrow 1, N_j$ do
            Randomly pick $\tilde{\mathcal{I}}_{k-1} \subset \{1, ..., n\}$ with $\mid \tilde{\mathcal{I}}_{k-1} \mid = b_j$
            $\nu_{k-1}^{(j)} \leftarrow \nabla f_{\tilde{\mathcal{I}}_{k-1}}(\tilde{w}_{k-1}^{(j)}) - \nabla f_{\tilde{\mathcal{I}}_{k-1}}(\tilde{w}_0^{(j)}) + g_j$
            $w_k^{(j)} \leftarrow w_{k-1}^{(j)} - \eta_j \nu_{k-1}^{(j)}$
        end for
        $\tilde{w}_j \leftarrow w_{N_j}^{(j)}$
    end for
    Output : (Smooth case) Sample $\tilde{x}_T^*$ from $(\tilde{x}_j)_{j=1}^T$ with $P(\tilde{x}_T^* = \tilde{x}_j) \propto$
  $\eta_j B_j / b_j$; (P-L case) $\tilde{x}_T$
end procedure
```

Each component $f_i(w)$ is possibly non-convex, $\mathbf{T}$ is the number of stages, $\tilde{w}_0$ is the initial point, $\mathbf{B} = (B_j)_{j=1}^T$ the batch sizes, $\mathbf{b} = (b_j)_{j=1}^T$ the mini-batch sizes, and $\eta = (\eta_j)_{j=1}^T$ the step sizes.

2 Parameter tuning

The introduced algorithm has a few hyperparameters that needs to be tuned. Those parameters are : the step sizes $\eta = (\eta_j)_{j=1}^T$, the batch sizes $\mathbf{B} = (B_j)_{j=1}^T$ and the mini-batch sizes $\eta = (\eta_j)_{j=1}^T$. Tuning hyperparameters is known to be costly but, fortunately, three default scheme with theoretical support was prodived by the autors. We would be using those default scheme in our experiments.

3 Convergence analysis

In this paper, the convergence analysis is conducted under different scenarios and hypothesis and the SCSG always obtained better theoretical results compared to SGD. But, does those results hold in practice ? Let's see !

4 Our results

Please refer to the notebook attached to this report for more details on our experiments and results.

4.1 The convex case

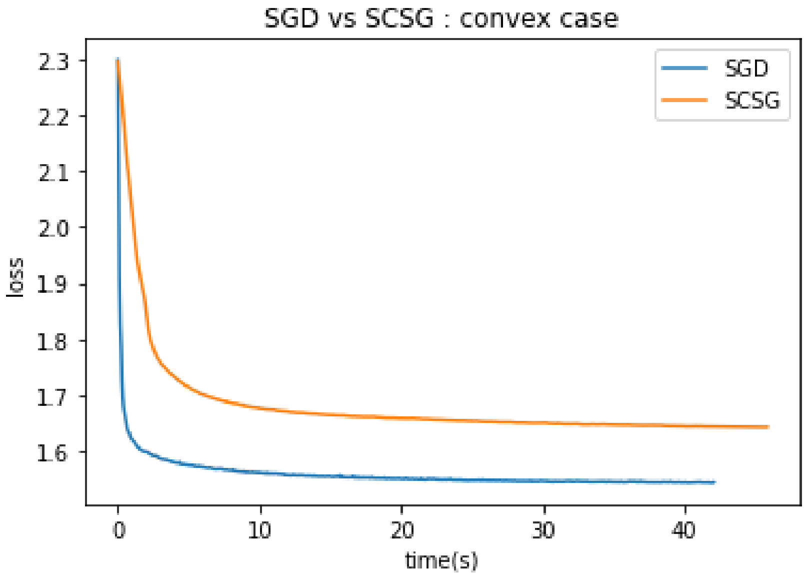

Figure 1: The convex case: convergence is very stable and fast.

4.2 The non convex case.

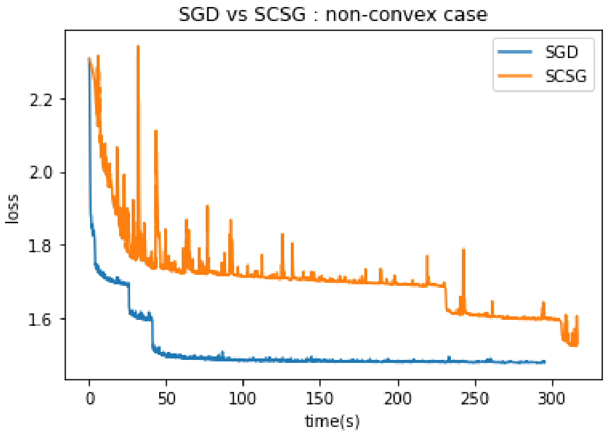

Figure 2: The non-convex case: convergence is less stable and is relatively slow.

Please refer to the notebook attached to this report for more details on our experiments and results.

References

[Bottou, 2010] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent.

[L. Lei, 2017] L. Lei, C. Ju, J. C. M. J. (2017). Non-convex finite-sum optimization via scsg methods.

[L. Lei, 2016] L. Lei, J. Chen, M. J. (2016). Less than a single pass: Stochastically controlled stochastic gradient method.