A TEMPLATE FOR THE *arxiv* STYLE

A PREPRINT

Igor Melnikov
Moscow Institute of Physics and Technology
Dolgoprudny, Russia
melnikov.ia@phystech.edu

Rustem Islamov
Institut Polytechnique de Paris
Palaiseau, France
rustem.islamov@ip-paris.fr

ABSTRACT

We analyse Newton-type methods of Empirical Risk Minimization problem for some Machine Learning model using Newton-type method accessing one data point per iteration. Specifically, we plan to improve stochastic variant of Newton's method. Unlike most other stochastic variants of second order methods, which require the evaluation of a large number of gradients and/or Hessians in each iteration to guarantee convergence, this methods do not have this shortcoming. We try to improve the performance of the algorithm by applying existing sampling strategies and incremental methods.

1 Introduction

Assume we have n training points (a_i, b_i) for $i \in \overline{1,n}$. We also assume n to be large. Let $f_i(x)$ be a loss function on i-th training point . We analyze second order methods solving Empirical Risk Minimization problem of the form.

$$\min_{x \in R} \left[f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right]. \tag{1}$$

As n is large this problem is typically solved by Stochastic Gradient Descent. First-order methods are well studied, and there are many papers describing them, such as (1). Stochastic Gradient Descen have cheap iterations independent of n, but with constant-stepsize due to high variance of the stochastic gradient has low accuracy convergence. Radius of this convergence is proportional to the variance of the stochastic gradient. Variance-reduced methods (2) have better convergence, however required number of iterations depends on the condition number.

Second-order methods, adapt to the curvature of the problem and thereby decrease their dependence on the condition number. The vanilla deterministic Newton method describes the following way:

$$x^{k+1} = x^k - \left(\nabla^2 f\left(x^k\right)\right)^{-1} \nabla f\left(x^k\right) \tag{2}$$

However this method does not directly translate to the stochastic case as convergence to the optimal of sample of functions is too quick. Despite first-order methods, this methods are poorly understood, a there are just a few researches connected to it (3). The problem is usually solved by large batch size. Nevertheless, by applying different sampling strategies we will smooth the step direction using the memory of past directions.

$$\begin{aligned} x^{k+1} = x^k - \left(\nabla^2 f\left(x^k\right)\right)^{-1} \nabla f\left(x^k\right) = \\ \left(\nabla^2 f\left(x^k\right)\right)^{-1} \left(\nabla^2 f\left(x^k\right) x^k - \nabla f\left(x^k\right)\right) = \\ \left(\sum_{i=1}^{n} \nabla^2 f_i\left(x^k\right)\right)^{-1} \sum_{i=1}^{n} \left(\nabla^2 f_i\left(x^k\right) x^k - \nabla f_i\left(x^k\right)\right) \end{aligned} \tag{3}$$

We will modify the algorithm presented by Dmitry Kovalev and Konstantin Mishchenko (4). We will be iteratively by t from 1 to $T \leq n$ construct a vector $w^t \in R^n$ by the following rules:

$$w_i^0 = x^0, \text{ for } i \in \overline{1,n}.$$
$$w_i^{t+1} = \begin{cases} w_i^t & i \notin S^t \\ x^{t+1} & i \in S^t \end{cases}$$

S^k is uniformly chosen random subset of $\{1, \ldots, n\}$ with size τ.

The algorithms looks the following way:

$$x^{k+1} = \left(\sum_{i=1}^{n} \nabla^2 f_i \left(w_i^k \right) \right)^{-1} \sum_{i=1}^{n} \left(\nabla^2 f_i \left(x^k \right) w_i^k - \nabla f_i \left(w_i^k \right) \right) \tag{4}$$

Often other sampling strategies may work better, so we plan to implement them.

References

[1] Ahmed Khaled, Othmane Sebbouh, Nicolas Loizou, Robert M. Gower, Peter Richtárik - "Unified Analysis of Stochastic Gradient Methods for Composite Convex and Smooth Optimization"

[2] Eduard Gorbunov, Filip Hanzely, Peter Richtarik - "A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent"

[3] Anton Rodomanov, Dmitry Kropotov - "A Superlinearly-Convergent Proximal Newton-type Method for the Optimization of Finite Sums"

[4] Dmitry Kovalev, Konstantin Mishchenko - "Stochastic Newton and Cubic Newton Methods with Simple Local Linear-Quadratic Rates"