Loopless stochastic methods. Overview

Dmitry Beresnev, d.beresnev@innopolis.university
Vsevolod Klyushev, v.klyushev@innopolis.university

Fall 2023

1 L-SVRG and L-Katyusha

This section is devoted to the article "Don't Jump Through Hoops and Remove Those Loops: SVRG and Katyusha are Better Without the Outer Loop".

1.1 Background

Authors deal with the **empirical risk minimization (finite-sum) problems**, which have the following form:

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

Finite-sum problems form the **dominant paradigm for training the supervised machine learning models**: $f_i(x)$ is the loss of the model x on data point i.

The most remarkable algorithms for solving (1) are *variance-reduced* stochastic gradients algorithms, which are significantly faster than SGD on (strongly) convex. However, these methods are not quite successful in general non-convex problems.

Two of the most notable variance-reduced algorithms are **SVRG** (stochastic variance-reduced gradient) method and its accelerated variant known as **Katyusha**. Katyusha accelerates SVRG via employment "negative momentum" idea. Both methods have **a double loop design**: at the beginning of the outer loop, , a full pass over the training data is made to compute **the total gradient** of f at r eference point w_k. In SVRG the reference point is chosen as the freshest iterate, and in Katyusha w_k is a weighted average of recent iterates. The total gradient is then used in the inner loop to *adjust* the stochastic gradient $\nabla f_i(x^k)$ (x^k is the current iterate). In particular, both methods perform the following adjustment:

$$g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k). \tag{2}$$

It turns out that as methods progress, the variance of g^k progressively decreases to zero, what effects in **significantly faster convergence**.

Under assumptions of L-smoothness and μ-strongly convexity of f, the iterations complexities are the followings:

- $\mathcal{O}((n+\frac{L}{\mu})\log\frac{1}{\epsilon})$ for SVRG
- $\mathcal{O}((n+\sqrt{\frac{nL}{\mu}})\log\frac{1}{\epsilon})$ (accelerated) for Katyusha

what is vast improvement on linear rate of GD and sublinear rate of SGD.

1.2 Problem statement

As explained below, the key trade-mark structural feature of SVRG and Katyusha is presence of the outer loop where a pass over full data is made. However, the outer loop causes such problems as

- The methods are hard to analyze
- One needs to decide at which point to terminate the inner loop and start the outer loop

Elaborating the second issue, the theoretically optimal inner loop size for SVRG depends on both L and μ. However, μ is not always known, and even if estimate is available, it can be very loose. Due to these issues, inner loop size is often chosen in a suboptimal way.

1.3 Main idea

Authors address the above issues by developing **loopless** variant of both SVRG and Katyusha: L-SVRG and L-Katyusha respectively. In these methods, authors remove the outer loop and replace it with a *biased coin-flip*, which on every iteration decides if the gradient $\nabla f(w^k)$ should be calculated. In particular, at each step the following happens:

- With small probability $p > 0$, the full pass over data is performed and the reference gradient $\nabla f(w^k)$ is updated
- With probability $1 - p$, the previous reference gradient is kept.

This procedure can also be interpreted as **having an outer loop of random length**.

1.4 Results

The paper demonstrates that both proposed methods, L-SVRG and L-Katyusha, have the following advantages:

1. Loopless methods are **easier to write down and analyze** than original double loop implementations
2. Loopless methods have the **same theoretical convergence rates**
3. Loopless methods are **superior to their loopy variants**
4. L-SVRG is extremely robust to the choice of p within the theoretical optimal interval

The first point follows directly from the article: all the necessary proofs, lemmas and theorems are quite simple and compact. Moreover, most other theoretical results, as optimal interval of p for L-SVRG, are not sophisticated either.

The second point is proved inside the article with the help of several lemmas and a couple of theorems.

The third and fourth points are demonstrated through numerical experiments. In particular, authors show the following results:

- *The performance of L-Katyusha is at least as good as that of Katyusha, and can be significantly faster in some causes*
- *Even the worst case for L-SVRG outperforms the best case for SVRG*

2 PAGE

This section is devoted to the article "PAGE: A simple and OPtimal Probabilistic Gradient Estimator for Nonconvex Optimization".

2.1 Problem statement

Authors deal with the following general optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{3}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is differentiable and possibly non-convex function.

Authors interested in function having **finite-sum** form

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{4}$$

where the functions f_i are also differentiable and possibly non-convex. Form (4) represents empirical risk optimization problems in machine learning. Moreover if the number of data samples n is very large or even infinite, then $f(x)$ is usually modeled via the **online** form

$$f(x) := \mathbb{E}_{\zeta\sim\mathcal{D}}[F(x,\zeta)], \tag{5}$$

for which solution is also applicable by letting $f_i(x) := F(x\zeta)$.

2.2 Background

There are several methods for (3) (e.g. SPIDER, SpiderBoost, SARAH, SSRGD) with gradient complexity:

- $O(n+\frac{\sqrt{n}}{\epsilon^2})$ and $\Omega(\frac{\sqrt{n}}{\epsilon^2})$ if $n\le O(\frac{1}{\epsilon^4})$ in the finite sum regime
- $O(b+\frac{\sqrt{b}}{\epsilon^2})$ in online regime

At the same time SVRG has gradient complexity $O(n+\frac{n^{2/3}}{\epsilon^2})$ in finite-sum regime and $\tilde{O}(b+\frac{\sqrt{b}}{\epsilon^2})$ for online case. These methods are complicated, often with double loop structure, and reliance on several hyperparameters. Moreover, there is no tight lower bound to show optimality of optimal methods in the online regime.

2.3 Main idea

PAGE method is based on vanilla SGD: in each iteration, Page uses the vanilla minibatch SGD update with probability p_t or ruses the previous gradient with a small adjustment, at a much lower computational cost, with probability $1-p_t$.

PAGE has following gradient complexities:

- $O(n+\frac{\sqrt{n}}{\epsilon^2})$ and $\Omega(n+\frac{\sqrt{n}}{\epsilon^2})$ in the finite sum regime
- $O(b+\frac{\sqrt{b}}{\epsilon^2})$ and $\Omega(b+\frac{\sqrt{b}}{\epsilon^2})$, where $b:=\min\{\frac{\sigma^2}{\epsilon^2},n\}$ in online regime

However, Page has faster linear convergence rates for nonconvex functions under Polyak-Lojasiewicz (PL) condition:

A function $f:\mathbb{R}^d\to\mathbb{R}$ satisfies PL condition if $\exists\mu>0$ such that

$$\|\nabla f(x)\|^2 \ge 2\mu(f(x)-f^*), \forall x\in\mathbb{R}^d.$$

PAGE has following gradient complexities under PL condition:

- $O((n+\sqrt{n}\kappa)\log\frac{1}{\epsilon})$ in the finite sum regime
- $O((b+\sqrt{b}\kappa)\log\frac{1}{\epsilon})$ in online regime

2.4 Experiments

Authors conduct several deep learning experiments for multi-class image classification. They compare PAGE algorithm with vanilla SGD by running standard LeNet, VGG, and ResNet models on MNIST and CIFAR-10 datasets. Results of the experiments show practical superiority that PAGE.

2.5 Results

- PAGE has tight lower bounds for both non-convex finite-sum and online optimization problems.
- PAGE optimal convergence results match lower bounds for both non-convex finite-sum and online problems.
- PAGE is simple and optimal algorithm for both non-convex finite-sum and online optimization.
- PAGE is easy to implement.
- PAGE can automatically switch to a faster linear convergence rate for nonconvex functions which satisfies PL condition.
- Experiments results confirm practical superiority of PAGE.