Momentum-Based Variance Reduction in Non-Convex SGD

Yanjie Ze, June 2021

0 Introduction

1 Motivation

two potential issues of **SVRG**:

1. Non-adaptive learning rates
2. Reliance on giant batch sizes to construct variance reduced gradients throughout the use of low-noise gradients calculated at a "checkpoint"

In this paper, we address both of these issues.

Present a new algorithm called **STOchastic Recursive Momemtum**.

Affect: Achieve variance reduction through the use of a variant of the momentum term.

SAG:

SVRG:

Procedure SVRG

Parameters update frequency m and learning rate η
Initialize $\tilde{w}_0$
Iterate: for $s = 1, 2, \ldots$
  $\tilde{w} = \tilde{w}_{s-1}$
  $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} \nabla\psi_i(\tilde{w})$
  $w_0 = \tilde{w}$
  Iterate: for $t = 1, 2, \ldots, m$
    Randomly pick $i_t \in \{1, \ldots, n\}$ and update weight
      $w_t = w_{t-1} - \eta(\nabla\psi_{i_t}(w_{t-1}) - \nabla\psi_{i_t}(\tilde{w}) + \tilde{\mu})$
  end
  option I: set $\tilde{w}_s = w_m$
  option II: set $\tilde{w}_s = w_t$ for randomly chosen $t \in \{0, \ldots, m-1\}$
end

Figure 1: Stochastic Variance Reduced Gradient

我的个人理解：

SVRG的缺陷主要在于两点。第一，giant batch。第二，learning rate是固定的。

2 Setting

We can access a stream of independent random variables:

$$\xi_1, \ldots, \xi_T \in \Xi$$

A sample function f that satisfies:

$$\forall t, \mathbf{x},\ \mathbb{E}[f(\mathbf{x}, \xi_t)|\mathbf{x}] = F(\mathbf{x})$$

Where $F(x)$ is the oracle function we can not access directly.

The noise of the gradients is bounded by σ^2:

$$\mathbb{E}[||\nabla f(\mathbf{x}, \xi_t) - \nabla F(\mathbf{x})||^2] \leq \sigma^2$$

Define:

$$F^\star = \inf_x F(\mathbf{x})$$

$$F^\star > -\infty$$

Assume our function f is L-smooth and G-Lipschitz:

$$\forall x, ||\nabla f(\mathbf{x})|| \leq G$$

$$\forall x \; and \; y, ||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})|| \leq L||\mathbf{x} - \mathbf{y}||$$

3 Notation

Gradient direction:

$$\mathbf{d_t} = (1-a)\mathbf{d_{t-1}} + a\nabla f(\mathbf{x_t}, \xi_t) + (1-a)(\nabla f(\mathbf{x_t}, \xi_t) - \nabla f(\mathbf{x_{t-1}}, \xi_\mathbf{t}))$$

Update formula:

$$\mathbf{x_{t+1}} = \mathbf{x_t} - \eta \mathbf{d_t}$$

Error term:

$$\epsilon_\mathbf{t} = \mathbf{d_t} - \nabla F(\mathbf{x_t})$$

Variables in Theorem 1:

$$k = \frac{bG^{\frac{2}{3}}}{L}$$

$$c = 28L^2 + G^2/(7Lk^3) = L^2(28 + 1/(7b^3))$$

$$w = max((4Lk)^3, 2G^2, (\frac{ck}{4L})^3) = G^2 max((4b)^3, 2, (28b + \frac{1}{7b^2})^3/64)$$

$$M = \frac{8}{k}(F(\mathbf{x_1}) - F^\star) + \frac{w^{1/3}\sigma^2}{4L^2k^2} + \frac{k^2c^2}{2L^2}\ln(T+2)$$

Variables in Algorithm STORM:

$$\eta_t \leftarrow \frac{k}{(w + \sum_{i=1}^t G_t^2)^{\frac{1}{3}}}$$

$$a_{t+1} \leftarrow c\eta_t^2$$

$$G_{t+1} \leftarrow ||\nabla f(\mathbf{x_{t+1}}, \eta_{\mathbf{t+1}})||$$

$$\mathbf{d_{t+1}} \leftarrow \nabla f(\mathbf{x_{t+1}}, \xi_{t+1}) + (1 - a_{t+1})(\mathbf{d}_t - \nabla f(\mathbf{x}_t, \xi_{t+1}))$$

4 Background: Momentum and Variance Reduction

$$\mathbf{d_t} = (1-a)\mathbf{d_{t-1}} + a\nabla f(\mathbf{x_t}, \xi_t)$$

$$\mathbf{x_{t+1}} = \mathbf{x_t} - \eta \mathbf{d_t}$$

Where a is small, i.e. $a = 0.1$

However, it's still unclear if the actual convergence rate can be improved by the momentum.

Hence, instead of showing that momentum in SGD works in the same way as in the noiseless case, we show that **a variant of momentum can provably reduce the variance of the gradients**.

$$\mathbf{d_t} = (1-a)\mathbf{d_{t-1}} + a\nabla f(\mathbf{x_t}, \xi_t) + (1-a)(\nabla f(\mathbf{x_t}, \xi_t) - \nabla f(\mathbf{x_{t-1}}, \xi_{\mathbf{t}}))$$

$$\mathbf{x_{t+1}} = \mathbf{x_t} - \eta \mathbf{d_t}$$

The only difference is a new term:

$$(1-a)(\nabla f(\mathbf{x_t}, \xi_t) - \nabla f(\mathbf{x_{t-1}}, \xi_{\mathbf{t}}))$$

5 Algorithm: Storm

Algorithm 1 STORM: STOchastic Recursive Momentum

Input: Parameters k, w, c, initial point $\boldsymbol{x}_1$
Sample ξ_1
$G_1 \leftarrow \|\nabla f(\boldsymbol{x}_1, \xi_1)\|$
$\boldsymbol{d}_1 \leftarrow \nabla f(\boldsymbol{x}_1, \xi_1)$
$\eta_0 \leftarrow \frac{k}{w^{1/3}}$
for $t = 1$ **to** T **do**
  $\eta_t \leftarrow \frac{k}{(w+\sum_{i=1}^{t} G_t^2)^{1/3}}$
  $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \eta_t \boldsymbol{d}_t$
  $a_{t+1} \leftarrow c\eta_t^2$
  Sample ξ_{t+1}
  $G_{t+1} \leftarrow \|\nabla f(\boldsymbol{x}_{t+1}, \xi_{t+1})\|$
  $\boldsymbol{d}_{t+1} \leftarrow \nabla f(\boldsymbol{x}_{t+1}, \xi_{t+1}) + (1 - a_{t+1})(\boldsymbol{d}_t - \nabla f(\boldsymbol{x}_t, \xi_{t+1}))$
end for
Choose $\hat{\boldsymbol{x}}$ uniformly at random from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$. (In practice, set $\hat{\boldsymbol{x}} = \boldsymbol{x}_T$).
return $\hat{\boldsymbol{x}}$

5 Theorem 1

Theorem 1. *Under the assumptions in Section 3, for any $b > 0$, we write $k = \frac{bG^{\frac{2}{3}}}{L}$. Set $c = 28L^2 + G^2/(7Lk^3) = L^2(28 + 1/(7b^3))$ and $w = \max\left((4Lk)^3, 2G^2, \left(\frac{ck}{4L}\right)^3\right) = G^2 \max\left((4b)^3, 2, (28b + \frac{1}{7b^2})^3/64\right)$. Then,* STORM *satisfies*

$$\mathbb{E}\left[\|\nabla F(\hat{\boldsymbol{x}})\|\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\boldsymbol{x}_t)\|\right] \leq \frac{w^{1/6}\sqrt{2M} + 2M^{3/4}}{\sqrt{T}} + \frac{2\sigma^{1/3}}{T^{1/3}},$$

where $M = \frac{8}{k}(F(\boldsymbol{x}_1) - F^\star) + \frac{w^{1/3}\sigma^2}{4L^2k^2} + \frac{k^2c^2}{2L^2}\ln(T+2)$.

$$k = \frac{bG^{\frac{2}{3}}}{L}$$

$$c = 28L^2 + G^2/(7Lk^3) = L^2(28 + 1/(7b^3))$$

$$w = max((4Lk)^3, 2G^2, (\frac{ck}{4L})^3) = G^2 max((4b)^3, 2, (28b + \frac{1}{7b^2})^3/64)$$

$$M = \frac{8}{k}(F(\mathbf{x_1}) - F^\star) + \frac{w^{1/3}\sigma^2}{4L^2k^2} + \frac{k^2c^2}{2L^2}\ln(T+2)$$

Explanation:

If there is no noise, which means $\sigma = 0$, then convergence rate is:

$$O(\frac{\ln T}{\sqrt{T}})$$

If there is noise (SGD), which means $\sigma \neq 0$, then convergence rate is:

$$O(\frac{2\sigma^{1/3}}{T^{1/3}})$$

In SGD case, this matches the optimal rate, which was obtained by SVRG-based algorithms that require a **mega batch**.

注意到，在第一项中，当G趋于0时，k趋于0，M趋于无穷，似乎第一项是趋于无穷的。但是，并不是这样。根据G-Lipschitz条件可得：

$$F(\mathbf{x_1}) - F^\star = O(G) \; and \; \sigma = O(G)$$

因此the numerators of M actually go to zero at least as fast as the denominator

注意到，当L=0时，no critical point，因为gradient都是相同的。

总的来说，M可以看作是一个$O(\log T)$的项。

6 Lyapunov potential function

In the theory of ordinary differential equations (ODEs), **Lyapunov functions** are scalar functions that may be used to prove the stability of an equilibrium of an ODE.

typical form:

$$\Phi_t = F(\mathbf{x_t})$$

Our form:

$$\Phi_t = F(\mathbf{x_t}) + z_t ||\epsilon_t||^2$$

Where $z_t \propto \eta_{t-1}^{-1}$ and ϵ is the error term.

7 Proof of Theorem 1

First we introduce several lemmas.

Lemma 1. *Suppose $\eta_t \le \frac{1}{4L}$ for all t. Then*

$$\mathbb{E}[F(\boldsymbol{x}_{t+1}) - F(\boldsymbol{x}_t)] \le \mathbb{E}\left[-\eta_t/4\|\nabla F(\boldsymbol{x}_t)\|^2 + 3\eta_t/4\|\boldsymbol{\epsilon}_t\|^2\right] .$$

Lemma 2. *With the notation in Algorithm 1, we have*

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^2/\eta_{t-1}\right] \le \mathbb{E}\left[2c^2\eta_{t-1}^3 G_t^2 + (1-a_t)^2(1+4L^2\eta_{t-1}^2)\|\boldsymbol{\epsilon}_{t-1}\|^2/\eta_{t-1} + 4(1-a_t)^2 L^2 \eta_{t-1}\|\nabla F(\boldsymbol{x}_{t-1})\|^2\right] .$$

Lemma 4. *Let $a_0 > 0$ and $a_1, \dots, a_T \ge 0$. Then*

$$\sum_{t=1}^{T} \frac{a_t}{a_0 + \sum_{i=1}^t a_i} \le \ln\left(1 + \frac{\sum_{i=1}^t a_i}{a_0}\right) .$$

Consider a Lyapunov function of the form:

$$\Phi_t = F(\mathbf{x_t}) + \frac{1}{32L^2\eta_{t-1}}||\epsilon_t||^2$$

We will upper bound $\Phi_{t+1} - \Phi_t$ for each t, which will allow us to bound Φ_T in terms of Φ_1 by summing over t.

$\mathbb{E}[\eta_t^{-1}||\epsilon_{\mathbf{t+1}}||^2 - \eta_{t-1}^{-1}||\epsilon_{\mathbf{t}}||^{\mathbf{2}}]$

Use Lemma 2, we first consider $\mathbb{E}[\eta_t^{-1}||\epsilon_{\mathbf{t+1}}||^2 - \eta_{t-1}^{-1}||\epsilon_{\mathbf{t}}||^{\mathbf{2}}]$:

$$\begin{aligned}&\mathbb{E}[\eta_t^{-1}||\epsilon_{\mathbf{t+1}}||^2 - \eta_{t-1}^{-1}||\epsilon_{\mathbf{t}}||^{\mathbf{2}}]\\ \le&\mathbb{E}\left[2c^2\eta_t^3 G_{t+1}^2 + \left(\eta_t^{-1}(1-a_{t+1})(1+4L^2\eta_t^2) - \eta_{t-1}^{-1}\right)||\epsilon||^2 + 4L^2\eta_t||\nabla F(\mathbf{x_t})||^2\right]\end{aligned}$$

There are three terms in the right side, and we denote them as A_t, B_t, C_t.

$$A_t = 2c^2\eta_t^3 G_{t+1}^2$$

$$B_t = (\eta_t^{-1}(1-a_{t+1})(1+4L^2\eta_t^2) - \eta_{t-1}^{-1})||\epsilon||^2$$

$$C_t = 4L^2\eta_t||\nabla F(\mathbf{x_t})||^2$$

Then let us focus on these terms individually.

For A_t:

$$\sum_{t=1}^{T} A_t = \sum_{t=1}^{T} 2c^2\eta_t^3 G_{t+1}^2 \leq 2k^3c^2\ln(T+2) \ (using\ Lemma\ 4)$$

For B_t:

$$B_t \leq (\eta_t^{-1} - \eta_{t-1}^{-1} + \eta_t(4L^2 - c))||\epsilon_{\mathbf{t}}||^2$$

$$\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \leq \frac{G^2}{7Lk^3}\eta_t$$

$$\eta_t(4L^2 - c) \leq -24L^2\eta_t - G^2\eta_t/(7Lk^3)$$

$$Thus, B_t \leq -24L^2\eta_t||\epsilon_{\mathbf{t}}||^2$$

For C_t:

We haven't done something on C_t yet.

Putting all this together, we can get:

$$\frac{1}{32L^2}\sum_{t=1}^{T}\left(\frac{||\epsilon_{t+1}||^2}{\eta_t} - \frac{||\epsilon_t||^2}{\eta_{t-1}}\right) \leq \frac{k^3c^2}{16L^2}\ln(T+2) + \sum_{t=1}^{T}\left[\frac{\eta_t}{8}||\nabla F(x_t)||^2 - \frac{3\eta_t}{4}||\epsilon_t||^2\right]$$

$\mathbb{E}[\Phi_{t+1} - \Phi_t]$

Now we are ready to analyze the potential Φ_t.

Since $\eta_t \leq \frac{1}{4L}$, we can use Lemma 1 to obtain:

$$\mathbb{E}[\Phi_{t+1} - \Phi_t] \leq \mathbb{E}\left[-\frac{\eta_t}{4}||\nabla F(x_t)||^2 + \frac{3\eta_t}{4}||\epsilon_t||^2 + \frac{1}{32L^2\eta_t}||\epsilon_{t+1}||^2 - \frac{1}{32L^2\eta_{t-1}}||\epsilon_t||^2\right]$$

Summing over t and using the formula in the last part, we can get:

$$\mathbb{E}[\Phi_{T+1} - \Phi_1] \leq \mathbb{E}\left[\frac{k^3 c^2}{16L^2} ln(T+2) - \sum_{t=1}^{T} \frac{\eta_t}{8} ||\nabla F(x_t)||^2\right]$$

Reordering the terms, we have:

$$\mathbb{E}\left[\sum_{t=1}^{T} \eta_t ||\nabla F(x_t)||^2\right] \leq 8(F(x_1) - F^\star) + \frac{w^{\frac{1}{3}} \sigma^2}{(4L^2 k)} + \frac{k^3 c^2}{(2L^2)} \ln(T+2)$$

$\mathbb{E}\left[\sum_{t=1}^{T} ||\nabla F(x_t)||^2\right]$

Now, we relate $\mathbb{E}\left[\sum_{t=1}^{T} \eta_t ||\nabla F(x_t)||^2\right]$ to $\mathbb{E}\left[\sum_{t=1}^{T} ||\nabla F(x_t)||^2\right]$.

First, since η_t is decreasing,

$$\mathbb{E}\left[\sum_{t=1}^{T} \eta_t ||\nabla F(x_t)||^2\right] \geq \mathbb{E}\left[\eta_T \sum_{t=1}^{T} ||\nabla F(x_t)||^2\right]$$

Now, from Cauchy-Schwarz inequality, for any random variables A and B we have:

$$\mathbb{E}[A^2]\mathbb{E}[B^2] \geq \mathbb{E}[AB]^2$$

Hence, setting:

$$A = \sqrt{\eta_T \sum_{t=1}^{T-1} ||\nabla F(x_t)||^2}$$

$$B = \sqrt{\frac{1}{\eta_T}}$$

We obtain:

$$\mathbb{E}\left[\eta_T \sum_{t=1}^{T-1} ||\nabla F(x_t)||^2\right] \mathbb{E}\left[\frac{1}{\eta_T}\right] \geq \mathbb{E}\left[\sqrt{\sum_{t=1}^{T-1} ||\nabla F(x_t)||^2}\right]^2$$

To simplify the result, we set:

$$M = \frac{1}{k}\left[8(F(x_1) - F^\star) + \frac{w^{\frac{1}{3}}\sigma^2}{(4L^2k)} + \frac{k^3c^2}{(2L^2)}\ln(T+2)\right]$$

Then we get:

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T-1}||\nabla F(x_t)||^2}\right]^2 \le \mathbb{E}\left[M\left(w + \sum_{t=1}^{T} G_t^2\right)^{\frac{1}{3}}\right]$$

Define $\zeta = \nabla f(x_t, \xi_t) - \nabla F(x_t)$, so that:

$$\mathbb{E}[||\zeta_t||^2] \le \sigma^2$$

Then, we have:

$$G_t^2 = ||\nabla F(x_t) + \zeta_t||^2 \le 2||\nabla F(x_t)||^2 + 2||\zeta_t||^2$$

And another formula:

$$(a+b)^{\frac{1}{3}} \le a^{\frac{1}{3}} + b^{\frac{1}{3}}$$

Plug them in, we obtain:

$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T-1}||\nabla F(x_t)||^2}\right]^2 \le M(w + 2T\sigma^2)^{\frac{1}{3}} + 2^{\frac{1}{3}}M\left(\mathbb{E}\left[\sqrt{\sum_{t=1}^{T-1}||\nabla F(x_t)||^2}\right]\right)^{\frac{2}{3}}$$

To simplify this inequality, we define:

$$X = \sqrt{\sum_{t=1}^{T}||\nabla F(x_t)||^2}$$

Then the above can be written as:

$$(\mathbb{E}\,[X])^2 \le M(w + 2T\sigma^2)^{\frac{1}{3}} + 2^{\frac{1}{3}}M(\mathbb{E}\,[X])^{\frac{2}{3}}$$

This means that

either

$$(\mathbb{E}\,[X])^2 \le M(w + 2T\sigma^2)^{\frac{1}{3}}$$

or

$$(\mathbb{E}\,[X])^2 \leq 2^{\frac{1}{3}}\,M(\mathbb{E}\,[X])^{\frac{2}{3}}$$

Thus, we can solve $\mathbb{E}[X]$:

$$\mathbb{E}[X] \leq \sqrt{2M}(w + 2T\sigma^2)^{\frac{1}{6}} + 2M^{\frac{3}{4}}$$

By Cauchy-Schwarz, we have:

$$\sum_{t=1}^{T} ||\nabla F(x_t)||/T \leq X/\sqrt{T}$$

And also,

$$(a+b)^{\frac{1}{3}} \leq a^{\frac{1}{3}} + b^{\frac{1}{3}}$$

Thus:

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{||\nabla F(x_t)||}{T}\right] \leq \frac{w^{\frac{1}{6}}\sqrt{2M} + 2M^{\frac{3}{4}}}{\sqrt{T}} + \frac{2\sigma^{\frac{1}{3}}}{T^{\frac{1}{3}}}$$