

STORM

Momentum-based Variance Reduction in Non-Convex SGD

Yanjie Ze

July 19, 2021

SJTU CSE

1. Background
2. Motivation
3. Setting
4. Algorithm
5. Theorem
6. Notation
7. Proof of Theorem 1
8. Proof of Lemma 2

Background

Variance Reduction Methods

Common methods:

- SAG(Stochastic Average Gradient)
- SDCA(Stochastic Dual Coordinate Ascent)
- SVRG((Stochastic Variance Reduction Gradient)
-

SVRG(Stochastic Variance Reduction Gradient) is one classical method among them, which achieves geometry convergence.

Theorem 1. Consider SVRG in Figure 1 with option II. Assume that all ψ_i are convex and both (5) and (6) hold with $\gamma > 0$. Let $w_* = \arg \min_w P(w)$. Assume that m is sufficiently large so that

$$\alpha = \frac{1}{\gamma\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1,$$

then we have geometric convergence in expectation for SVRG:

$$\mathbb{E} P(\tilde{w}_s) \leq \mathbb{E} P(w_*) + \alpha^s [P(\tilde{w}_0) - P(w_*)]$$

Figure 1: Convergence of SVRG

Procedure SVRG

Parameters update frequency m and learning rate η

Initialize \tilde{w}_0

Iterate: for $s = 1, 2, \dots$

$$\tilde{w} = \tilde{w}_{s-1}$$

$$\tilde{\mu} = \frac{1}{n} \sum_{i=1}^n \nabla \psi_i(\tilde{w})$$

$$w_0 = \tilde{w}$$

Iterate: for $t = 1, 2, \dots, m$

Randomly pick $i_t \in \{1, \dots, n\}$ 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 $t \in \{0, \dots, m-1\}$

end

Figure 2: Stochastic Variance Reduction Gradient

Adaptive Learning Rate

Adaptive Learning Rate: choose the values η_t in some data-dependent way so as to reduce the need for tuning the values of η_t manually.

In the non-convex setting, adaptive learning rates can be shown to improve the convergence rate of SGD to

$$O\left(\frac{1}{\sqrt{T}} + \left(\frac{\sigma^2}{T}\right)^{\frac{1}{4}}\right)$$

Where σ^2 is a bound on the variance of $\nabla f(x_t)$.

Motivation

Motivation

- Based on Momentum:

$$\mathbf{d}_t = (1 - a)\mathbf{d}_{t-1} + a\nabla f(\mathbf{x}_t, \xi_t)$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\mathbf{d}_t$$

- Using adaptive learning rate,
which has not been used in **Variance Reduction Methods in the non-convex setting**(only one method in convex setting)
- Removing the need for 'giant batch'
Most Variance Reduction Methods require the calculation of gradients at checkpoints, such as SVRG.

Momentum:

$$\mathbf{d}_t = (1 - a)\mathbf{d}_{t-1} + a\nabla f(\mathbf{x}_t, \xi_t)$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\mathbf{d}_t$$

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.

Momentum-based

Hence, instead of showing that momentum in SGD works in the same way as in the noiseless case, this work shows that **a variant of momentum can provably reduce the variance of the gradients**, as shown below.

SVRG:

$$\mathbf{d}_t = (1 - a)\mathbf{d}_{t-1} + a\nabla f(\mathbf{x}_t, \xi_t) + (1 - a)(\nabla f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_{t-1}, \xi_t))$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\mathbf{d}_t$$

The only difference is a new term:

$$(1 - a)(\nabla f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_{t-1}, \xi_t))$$

Setting

We can access a stream of independent random variables:

$$\xi_1, \dots, \xi_T \in \Xi$$

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 :

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

Assume our function f is L -smooth and G -Lipschitz, which is to say:

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

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

Also, this work gives the convergence rate in the setting **without G-Lipschitz** in the appendix.

Algorithm

Algorithm 1 STORM: STOchastic Recursive Momentum

```

1: Input: Parameters  $k, w, c$ , initial point  $\mathbf{x}_1$ 
2: Sample  $\xi_1$ 
3:  $G_1 \leftarrow \|\nabla f(\mathbf{x}_1, \xi_1)\|$ 
4:  $\mathbf{d}_1 \leftarrow \nabla f(\mathbf{x}_1, \xi_1)$ 
5:  $\eta_0 \leftarrow \frac{k}{w^{1/3}}$ 
6: for  $t = 1$  to  $T$  do
7:    $\eta_t \leftarrow \frac{k}{(w + \sum_{i=1}^t G_i^2)^{1/3}}$ 
8:    $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta_t \mathbf{d}_t$ 
9:    $a_{t+1} \leftarrow c\eta_t^2$ 
10:  Sample  $\xi_{t+1}$ 
11:   $G_{t+1} \leftarrow \|\nabla f(\mathbf{x}_{t+1}, \xi_{t+1})\|$ 
12:   $\mathbf{d}_{t+1} \leftarrow \nabla f(\mathbf{x}_{t+1}, \xi_{t+1}) + (1 - a_{t+1})(\mathbf{d}_t - \nabla f(\mathbf{x}_t, \xi_{t+1}))$ 
13: end for
14: Choose  $\hat{\mathbf{x}}$  uniformly at random from  $\mathbf{x}_1, \dots, \mathbf{x}_T$ . (In practice, set  $\hat{\mathbf{x}} = \mathbf{x}_T$ ).
15: return  $\hat{\mathbf{x}}$ 

```

Figure 3: STOchastic Recursive Momentum

Intuition 1

The update of the gradient direction is:

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

This means the update gradient direction d_{t+1} is determined by:

- Current gradient (positive impact on d_{t+1})
- Last gradient (negative impact on d_{t+1})
- Accumulative gradient \mathbf{d}_t

Intuition 2

The update rate of \mathbf{x}_{t+1} is:

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

Which means as time goes, the rate decreases to 0.

Intuition 3

The update rate of d_{t+1} is:

$$a_{t+1} = c\eta_t^2 = c \cdot \frac{k^2}{(\omega + \sum_{i=1}^t G_i^2)^{\frac{2}{3}}}$$

Which means as time goes, the rate also goes to 0.

Theorem

Theorem 1

Theorem 1 gives the convergence rate of STORM with G-Lipschitz.

We will go through the proof of this theorem in detail.

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

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

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

Figure 4: Convergence of STORM with G-Lipschitz

Theorem 2

Theorem 2 gives the convergence rate of STORM without G-Lipschitz.

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

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

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

Figure 5: Convergence of STORM without G-Lipschitz

Notation

Gradient direction:

$$\mathbf{d}_t = (1 - a)\mathbf{d}_{t-1} + a\nabla f(\mathbf{x}_t, \xi_t) + (1 - a)(\nabla f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_{t-1}, \xi_t))$$

Update formula:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{d}_t$$

Error term:

$$\epsilon_t = \mathbf{d}_t - \nabla F(\mathbf{x}_t)$$

Variables in Theorem 1:

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

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

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

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

Variables in Algorithm STORM:

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

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

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

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

Proof of Theorem 1

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.

Lyapunov potential function

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 .

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

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

The following technical observation is key to our analysis of STORM: it provides a recurrence that enables us to bound the variance of the estimates \mathbf{d}_t .

Lemma 2. With the notation in Algorithm [7](#), we have

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

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

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

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

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

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

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

$$\begin{aligned} A_t &= 2c^2\eta_t^3 G_{t+1}^2 \\ B_t &= (\eta_t^{-1}(1 - a_{t+1})(1 + 4L^2\eta_t^2) - \eta_{t-1}^{-1})||\epsilon||^2 \\ C_t &= 4L^2\eta_t||\nabla F(\mathbf{x}_t)||^2 \end{aligned}$$

Then let us focus on these terms individually.

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

For A_t :

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

For B_t :

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

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

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

$$\text{Thus, } B_t \leq -24L^2 \eta_t \|\epsilon_t\|^2$$

For C_t :

We haven't done something on C_t yet.

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

Putting all this together, we can get:

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

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

Now we are ready to analyze the potential Φ_t .

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

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

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

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:

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

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

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,

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

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

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:

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

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

To simplify the result, we set:

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

Then we get:

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

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

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

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

Then, we have:

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

And another formula:

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

Plug them in, we obtain:

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

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

To simplify this inequality, we define:

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

Then the above can be written as:

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

This means that

either

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

or

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

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

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

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

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:

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

Proof of Lemma 2

$$\mathbb{E} [\|\epsilon_t\|^2/\eta_{t-1}] \leq \mathbb{E} [2c^2\eta_{t-1}^3 G_t^2 + (1-a_t)^2(1+4L^2\eta_{t-1}^2)\|\epsilon_{t-1}\|^2/\eta_{t-1} + 4(1-a_t)^2 L^2 \eta_{t-1} \|\nabla F(\mathbf{x}_{t-1})\|^2] .$$

Proof. First, observe that

$$\begin{aligned} & \mathbb{E} [\eta_{t-1}^3 \|\nabla f(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{x}_t)\|^2] \\ &= \mathbb{E} [\eta_{t-1}^3 (\|\nabla f(\mathbf{x}_t, \xi_t)\|^2 + \|\nabla F(\mathbf{x}_t)\|^2 - 2\nabla f(\mathbf{x}_t, \xi_t) \cdot \nabla F(\mathbf{x}_t))] \\ &= \mathbb{E} [\eta_{t-1}^3 \mathbb{E} [\|\nabla f(\mathbf{x}_t, \xi_t)\|^2 + \|\nabla F(\mathbf{x}_t)\|^2 - 2\nabla f(\mathbf{x}_t, \xi_t) \cdot \nabla F(\mathbf{x}_t) | \xi_1, \dots, \xi_{t-1}]] \\ &= \mathbb{E} [\eta_{t-1}^3 (\|\nabla f(\mathbf{x}_t, \xi_t)\|^2 - \|\nabla F(\mathbf{x}_t)\|^2)] \\ &\leq \mathbb{E} [\eta_{t-1}^3 \|\nabla f(\mathbf{x}_t, \xi_t)\|^2] . \end{aligned} \tag{5}$$

In the same way, we also have that

$$\begin{aligned} & \mathbb{E} [\eta_{t-1}^{-1} (1-a_t^2) \|\nabla f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_{t-1}, \xi_t) - \nabla F(\mathbf{x}_t) + \nabla F(\mathbf{x}_{t-1})\|^2] \\ &\leq \mathbb{E} [\eta_{t-1}^{-1} (1-a_t^2) \|\nabla f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_{t-1}, \xi_t)\|^2] . \end{aligned} \tag{6}$$

In (5), the third line is based on the property of the expectation.(?)

By definition of ϵ_t and the notation in Algorithm [1](#), we have $\epsilon_t = \mathbf{d}_t - \nabla F(\mathbf{x}_t) = \nabla f(\mathbf{x}_t, \xi_t) + (1 - a_t)(\mathbf{d}_{t-1} - \nabla f(\mathbf{x}_{t-1}, \xi_t)) - \nabla F(\mathbf{x}_t)$. Hence, we can write

$$\begin{aligned}
\mathbb{E} [\eta_{t-1}^{-1} \|\epsilon_t\|^2] &= \mathbb{E} [\eta_{t-1}^{-1} \|\nabla f(\mathbf{x}_t, \xi_t) + (1 - a_t)(\mathbf{d}_{t-1} - \nabla f(\mathbf{x}_{t-1}, \xi_t)) - \nabla F(\mathbf{x}_t)\|^2] \\
&= \mathbb{E} [\eta_{t-1}^{-1} \|a_t(\nabla f(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{x}_t)) + (1 - a_t)(\nabla f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_{t-1}, \xi_t) - \nabla F(\mathbf{x}_t) + \nabla F(\mathbf{x}_{t-1})) \\
&\quad + (1 - a_t)(\mathbf{d}_{t-1} - \nabla F(\mathbf{x}_{t-1}))\|^2] \\
&\leq \mathbb{E} [2c^2\eta_{t-1}^3 \|\nabla f(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{x}_t)\|^2 + 2\eta_{t-1}^{-1}(1 - a_t)^2 \|\nabla f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_{t-1}, \xi_t) - \nabla F(\mathbf{x}_t) + \nabla F(\mathbf{x}_{t-1})\|^2 \\
&\quad + \eta_{t-1}^{-1}(1 - a_t)^2 \|\epsilon_{t-1}\|^2] \\
&\leq \mathbb{E} [2c^2\eta_{t-1}^3 \|\nabla f(\mathbf{x}_t, \xi_t)\|^2 + 2\eta_{t-1}^{-1}(1 - a_t)^2 \|\nabla f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_{t-1}, \xi_t)\|^2 + \eta_{t-1}^{-1}(1 - a_t)^2 \|\epsilon_{t-1}\|^2] \\
&\leq \mathbb{E} [2c^2\eta_{t-1}^3 G_t^2 + 2\eta_{t-1}^{-1}(1 - a_t)^2 L^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \eta_{t-1}^{-1}(1 - a_t)^2 \|\epsilon_{t-1}\|^2] \\
&= \mathbb{E} [2c^2\eta_{t-1}^3 G_t^2 + 2(1 - a_t)^2 L^2 \eta_{t-1} \|\mathbf{d}_{t-1}\|^2 + \eta_{t-1}^{-1}(1 - a_t)^2 \|\epsilon_{t-1}\|^2] \\
&= \mathbb{E} [2c^2\eta_{t-1}^3 G_t^2 + 2(1 - a_t)^2 L^2 \eta_{t-1} \|\epsilon_{t-1} + \nabla F(\mathbf{x}_{t-1})\|^2 + \eta_{t-1}^{-1}(1 - a_t)^2 \|\epsilon_{t-1}\|^2] \\
&\leq \mathbb{E} [2c^2\eta_{t-1}^3 G_t^2 + 4(1 - a_t)^2 L^2 \eta_{t-1} (\|\epsilon_{t-1}\|^2 + \|\nabla F(\mathbf{x}_{t-1})\|^2) + \eta_{t-1}^{-1}(1 - a_t)^2 \|\epsilon_{t-1}\|^2] \\
&= \mathbb{E} [2c^2\eta_{t-1}^3 G_t^2 + \eta_{t-1}^{-1}(1 - a_t)^2 (1 + 4L^2\eta_{t-1}^2) \|\epsilon_{t-1}\|^2 + 4(1 - a_t)^2 L^2 \eta_{t-1} \|\nabla F(\mathbf{x}_{t-1})\|^2],
\end{aligned}$$

where in the first inequality we used Lemma [3](#) (See Appendix [A](#)) and $\|\mathbf{x} + \mathbf{y}\|^2 \leq 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$, in the second inequality we used [5](#) and [6](#), in the third one the Lipschitzness and smoothness of the functions f , and in the last inequality we used again $\|\mathbf{x} + \mathbf{y}\|^2 \leq 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$. \square

Thanks!