Adaptive Accelerated (Extra)-Gradient Methods with Variance Reduction

Zijian Liu*[1], Ta Duy Nguyen*[1], Alina Ene[1], Huy Le Nguyen[2]

(1) Boston University and (2) Northeastern University

Variance Reduction in Finite Sum Optimization

$$\min_{x\in\mathcal{X}} \left\{ \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\}, \quad f_i \text{ is } \beta\text{-smooth}$$

- Variance reduction methods, starting from SVRG (Johnson & Zhang, 2013) show significant improvement over classic methods
- Accelerated VR methods were first proposed in Katyusha (Allen-Zhu, 2017)
- State-of-the-art: VARAG (Lan et at., 2019): $O\left(n\log n + \sqrt{\frac{n\beta}{\epsilon}}\right)$ and VRADA (Song et al, 2020): $O\left(n\log\log n + \sqrt{\frac{n\beta}{\epsilon}}\right)$

  gradient evaluations.
- Drawback: β must be known to set the step size.

Our work: Two adaptive accelerated VR algorithms, not require knowing β, same convergence guarantee

AdaVRAE: Extra-Gradient Method

$A_0^{(s)} = A_T^{(s-1)} - T(a^{(s)})^2$

$z_0^{(s)} = z_T^{(s-1)},\ g_0^{(s)} = g_T^{(s-1)}$

Inner loop: for $t = 1..T := n$

$$A_t^{(s)} = A_{t-1}^{(s)} + a^{(s)} + (a^{(s)})^2$$

$$x_t^{(s)} = \arg\min_x \left\{ a^{(s)} \langle g_{t-1}^{(s)}, x \rangle + \frac{\gamma_{t-1}^{(s)}}{2} \|x - z_{t-1}^{(s)}\|^2 \right\}$$

$$\bar{x}_t^{(s)} = \frac{1}{A_t^{(s)}} \left(A_{t-1}^{(s)} \bar{x}_{t-1}^{(s)} + a^{(s)} x_t^{(s)} + (a^{(s)})^2 u^{(s-1)} \right)$$

$$g_t^{(s)} = \nabla f_i(\bar{x}_t^{(s)}) - \nabla f_i(u^{(s-1)}) + \nabla f(u^{(s-1)}),\ i \sim \text{Uniform}([n]), \quad \text{or} \quad g_t^{(s)} = \nabla f(\bar{x}_t^{(s)}),\ \text{if } t = T$$

$$\gamma_t^{(s)} = \frac{1}{\eta} \sqrt{\eta^2 (\gamma_{t-1}^{(s)})^2 + (a^{(s)})^2 \|g_t^{(s)} - g_{t-1}^{(s)}\|^2} \longrightarrow \gamma_0^{(s)} = \gamma_T^{(s-1)}$$

$$z_t^{(s)} = \arg\min_z \left\{ a^{(s)} \langle g_t^{(s)}, z \rangle + \frac{\gamma_{t-1}^{(s)}}{2} \|z - z_{t-1}^{(s)}\|^2 + \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2} \|z - x_t^{(s)}\|^2 \right\}$$

Update checkpoint: $u^{(s)} = \bar{x}_T^{(s)}$

AdaVRAE: Extra-Gradient Method

$$A_0^{(s)} = A_T^{(s-1)} - T(a^{(s)})^2$$

$$z_0^{(s)} = z_T^{(s-1)},\ g_0^{(s)} = g_T^{(s-1)}$$

Inner loop: for $t = 1..T := n$

$$A_t^{(s)} = A_{t-1}^{(s)} + a^{(s)} + (a^{(s)})^2$$

$$x_t^{(s)} = \arg\min_x \left\{ a^{(s)}\langle g_{t-1}^{(s)}, x\rangle + \frac{\gamma_{t-1}^{(s)}}{2}\|x - z_{t-1}^{(s)}\|^2 \right\}$$

$$\bar{x}_t^{(s)} = \frac{1}{A_t^{(s)}}\left(A_{t-1}^{(s)}\bar{x}_{t-1}^{(s)} + a^{(s)}x_t^{(s)} + (a^{(s)})^2 u^{(s-1)}\right)$$

$$g_t^{(s)} = \nabla f_i(\bar{x}_t^{(s)}) - \nabla f_i(u^{(s-1)}) + \nabla f(u^{(s-1)}),\ i \sim \text{Uniform}([n]), \quad \text{or} \quad g_t^{(s)} = \nabla f(\bar{x}_t^{(s)}),\ \text{if } t = T$$

$$\gamma_t^{(s)} = \frac{1}{\eta}\sqrt{\eta^2(\gamma_{t-1}^{(s)})^2 + (a^{(s)})^2\|g_t^{(s)} - g_{t-1}^{(s)}\|^2} \longrightarrow \gamma_0^{(s)} = \gamma_T^{(s-1)}$$

$$z_t^{(s)} = \arg\min_z \left\{ a^{(s)}\langle g_t^{(s)}, z\rangle + \frac{\gamma_{t-1}^{(s)}}{2}\|z - z_{t-1}^{(s)}\|^2 + \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\|z - x_t^{(s)}\|^2 \right\}$$

Update checkpoint: $u^{(s)} = \bar{x}_T^{(s)}$

Theorem:

Let $s_0 = \lceil \log_2 \log_2 4n \rceil$ and

$$a^{(s)} = \begin{cases} (4n)^{-\frac{1}{2^s}} & \text{if } s \le s_0 \\ \frac{s - s_0 + \frac{1}{2}}{3} & \text{otherwise} \end{cases}$$

Then

$$\#\text{grads} = O\left(n \log\log n + \sqrt{\frac{n\beta}{\epsilon}}\right)$$

AdaVRAG: Single Projection Gradient Method

$\bar{x}_0^{(s)} = a^{(s)} x_0^{(s)} + (1 - a^{(s)}) u^{(s-1)}$

Inner loop: for $t = 1..T := n$

$$g_t^{(s)} = \nabla f_i(\bar{x}_{t-1}^{(s)}) - \nabla f_i(u^{(s-1)}) + \nabla f(u^{(s-1)}),\ i \sim \text{Uniform}([n])$$

$$x_t^{(s)} = \arg\min_x \left\{ \langle g_t^{(s)}, x \rangle + \frac{\gamma_{t-1}^{(s)} q^{(s)}}{2} \|x - x_{t-1}^{(s)}\|^2 \right\} \longrightarrow x_0^{(s)} = x_T^{(s-1)}$$

$$\bar{x}_t^{(s)} = a^{(s)} x_t^{(s)} + (1 - a^{(s)}) u^{(s-1)}$$

$$\gamma_t^{(s)} = \gamma_{t-1}^{(s)} \sqrt{1 + \frac{\|x_t^{(s)} - x_{t-1}^{(s)}\|^2}{\eta^2}} \longrightarrow \gamma_0^{(s)} = \gamma_T^{(s-1)}$$

Update checkpoint: $u^{(s)} = \frac{1}{T} \sum_{t=1}^{T} \bar{x}_t^{(s)}$

AdaVRAG: Single Projection Gradient Method

$\bar{x}_0^{(s)} = a^{(s)} x_0^{(s)} + (1 - a^{(s)}) u^{(s-1)}$

Inner loop: for $t = 1..T := n$

$$g_t^{(s)} = \nabla f_i(\bar{x}_{t-1}^{(s)}) - \nabla f_i(u^{(s-1)}) + \nabla f(u^{(s-1)}),\ i \sim \text{Uniform}([n])$$

$$x_t^{(s)} = \arg\min_x \left\{ \langle g_t^{(s)}, x \rangle + \frac{\gamma_{t-1}^{(s)} q^{(s)}}{2} \|x - x_{t-1}^{(s)}\|^2 \right\} \longrightarrow x_0^{(s)} = x_T^{(s-1)}$$

$$\bar{x}_t^{(s)} = a^{(s)} x_t^{(s)} + (1 - a^{(s)}) u^{(s-1)}$$

$$\gamma_t^{(s)} = \gamma_{t-1}^{(s)} \sqrt{1 + \frac{\|x_t^{(s)} - x_{t-1}^{(s)}\|^2}{\eta^2}} \longrightarrow \gamma_0^{(s)} = \gamma_T^{(s-1)}$$

Update checkpoint: $u^{(s)} = \frac{1}{T} \sum_{t=1}^{T} \bar{x}_t^{(s)}$

Theorem:

Let $s_0 = \lceil \log_2 \log_2 4n \rceil$ and

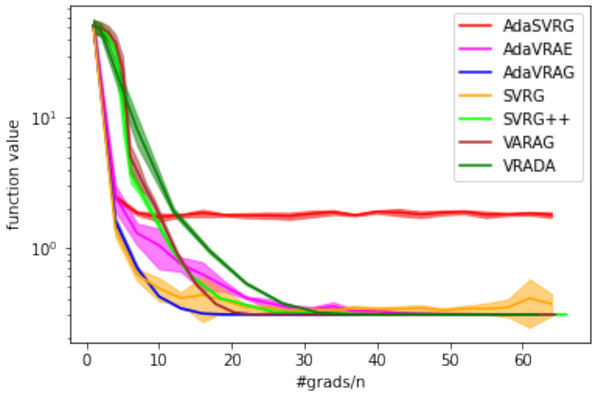
$$a^{(s)} = \begin{cases} 1 - (4n)^{-\frac{1}{2^s}} & \text{if } s \le s_0 \\ \frac{c}{s - s_0 + 2c}, & \text{otherwise} \end{cases},$$

$$q^{(s)} = \begin{cases} \frac{1}{(1 - a^{(s)}) a^{(s)}} & \text{if } s \le s_0 \\ \frac{8(2 - a^{(s)}) a^{(s)}}{3(1 - a^{(s)})} & \text{otherwise} \end{cases}$$

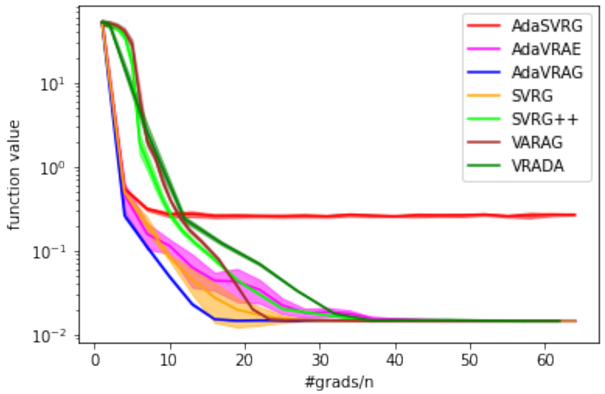
Then

$$\#\text{grads} = O\left(n \log\log n + \sqrt{\frac{n\beta \log \beta}{\epsilon}} \right)$$

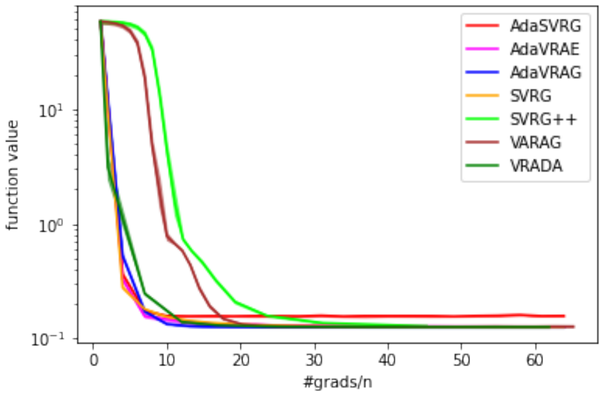
Experimental Results



a1a



mushroom



w8a

Test function values for logistic loss