Adaptive SVRG Methods under Error Bound Conditions with Unknown Growth Parameter

Yi Xu†, Qihang Lin‡, Tianbao Yang†

†Computer Science Department, ‡Management Sciences Department, The University of Iowa, Iowa City, IA, USA

Finite-sum Convex Problem

The optimization problem of interest:

$$\min_{x\in\Omega} F(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \Psi(x), \tag{1}$$

where $f_i(x)$ is convex and $\Psi(x)$ is proper, lower-semicontinuous and convex.
Let Ω_*, F_* denote the set of optimal solutions and the optimal value, respectively.

- We make the following assumptions:
  a. There exist $x_0 \in \Omega$ and $\epsilon_0 \ge 0$ s.t. $F(x_0) - F_* \le \epsilon_0$;
  b. Ω_* is a non-empty convex compact set;
  c. f_i is differential whose gradient is L_i-Lipschitz continuous, i.e. for all $x, y \in \Omega$,
  $$f_i(x) - f_i(y) \le \langle \nabla f_i(y), x - y\rangle + \frac{L_i}{2}\|x-y\|_2^2;$$
  d. $L \triangleq \max_i L_i$ is given or can be estimated for the problem.
- $f(x) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(x)$ is also continuously differential convex function whose gradient is L_f-Lipschitz continuous, where $L_f = \frac{1}{n}\sum_{i=1}^n L_i$.

Related Work

- Under the strong convexity of the objective function, stochastic variance reduced gradient (SVRG) method [1] and its proximal variant [2] achieve linear convergence.
- SVRG++ [3] can cope with non-strongly convex problems, however, it only has sublinear convergence (e.g., requiring a $O(1/\epsilon)$ iteration complexity to achieve an ϵ-optimal solution).
- Recent studies on optimization showed that leveraging the quadratic error bound (QEB) condition can open a new door to the linear convergence without strong convexity [4-9]
- The issue is that these methods (for example, SVRG) require to know the parameter c (analogous to the strong convexity parameter) in the QEB for setting the number of iterations of inner loops, which is usually unknown and difficult to estimate.

Hölderian error bound

Definition 1. *A function $F(x)$ is said to satisfy a* ***Hölderian error bound (HEB)*** *condition on a compact set Ω if there exist $\theta \in (0, 1/2]$ and $c > 0$ such that for any $x \in \Omega$*

$$\|x - x_*\|_2 \le c(F(x) - F_*)^{\theta}, \tag{2}$$

where x_* denotes the closest optimal solution to x.

- A special case of HEB is **quadratic error bound (QEB)**:
  $$\|x - x_*\|_2 \le c(F(x) - F(x_*))^{1/2}, \forall x \in \Omega, \tag{3}$$
  One example satisfying QEB is strongly convex function.
- The above inequality in HEB can always hold for $\theta = 0$ on a compact set Ω.
- If a HEB condition with $\theta \in (1/2, 1]$ holds, it can be reduced to the QEB condition provided that $F(x) - F_*$ is bounded over Ω.

SVRG under the HEB condition

Algorithm 1 SVRG under HEB (SVRG$^{\text{HEB}}(x_0, T_1, R, \theta)$)

Input: $x_0 \in \Omega$, number of inner initial iterations T_1, number of outer loops R.
$\bar{x}^{(0)} = x_0$
for $r = 1, 2, \ldots, R$ **do**
  $\bar{g}_r = \nabla f(\bar{x}^{(r-1)})$, $x_0^{(r)} = \bar{x}^{(r-1)}$
  for $t = 1, 2, \ldots, T_r$ **do**
    Choose $i_t \in \{1, \ldots, n\}$ uniformly at random.
    $g_t^{(r)} = \nabla f_{i_t}(x_{t-1}^{(r)}) - \nabla f_{i_t}(\bar{x}^{(r-1)}) + \bar{g}_r$.
    $x_t^{(r)} = \arg\min_{x\in\Omega}\langle g_t^{(r)}, x - x_{t-1}^{(r)}\rangle + \frac{1}{2\eta}\|x - x_{t-1}^{(r)}\|_2^2 + \Psi(x)$.
  end for
  $\bar{x}^{(r)} = \frac{1}{T_r}\sum_{t=1}^{T_r} x_t^{(r)}$, $T_{r+1} = 2^{1-2\theta}T_r$
end for
Output: $\bar{x}^{(R)}$

Theorem 1. *Assume problem (1) satisfies the HEB condition with $\theta \in (0, 1/2]$. Let $\eta = 1/(36L)$, and $T_1 \ge 81Lc^2(1/\epsilon_0)^{1-2\theta}$ (T_1* ***depends on*** *c). By running SVRG$^{\text{HEB}}$ with $R = \lceil \log_2 \frac{\epsilon_0}{\epsilon}\rceil$, we have $\mathrm{E}[F(\bar{x}^{(R)}) - F_*] \le \epsilon$. The iteration complexity of SVRG$^{\text{HEB}}$ in expectation is $O(n\log(\epsilon_0/\epsilon) + Lc^2\max\{\frac{1}{\epsilon^{1-2\theta}}, \log(\epsilon_0/\epsilon)\})$.*

- when $\theta = 1/2$ (i.e, the QEB condition holds), Algorithm 1 reduces to the standard SVRG method under strong convexity, and the iteration complexity becomes $O((n + Lc^2)\log(\epsilon_0/\epsilon))$, which is the same as that of the standard SVRG with Lc^2 mimicking the condition number of the problem.
- when $\theta = 0$ (i.e., with only the smoothness assumption), Algorithm 1 reduces to SVRG++ with one difference, where in SVRG$^{\text{HEB}}$ the initial point and the reference point for each outer loop are the same but are different in SVRG++, and the iteration complexity of SVRG$^{\text{HEB}}$ becomes $O(n\log(\epsilon_0/\epsilon) + \frac{Lc^2}{\epsilon})$ that is similar to that of SVRG++.
- for intermediate $\theta \in (0, 1/2)$, a faster convergence than SVRG++ can be obtained.

Adaptive SVRG for $\theta \in (0, 1/2)$

Algorithm 2 SVRG under HEB with Restarting: SVRG$^{\text{HEB-RS}}$

Input: $x^{(0)} \in \Omega$, a small value $c_0 > 0$, and $\theta \in (0, 1/2)$.
Initialization: $T_1^{(1)} = 81Lc_0^2(1/\epsilon_0)^{1-2\theta}$
for $s = 1, 2, \ldots, S$ **do**
  $x^{(s)}$=SVRG$^{\text{HEB}}$ $(x^{(s-1)}, T_1^{(s)}, R, \theta)$
  $T_1^{(s+1)} = 2^{1-2\theta}T_1^{(s)}$
end for

Main Result 1

Theorem 2. Assume problem (1) satisfies the HEB with $\theta \in (0, 1/2)$. Let $c_0 \le c$, $\epsilon \le \frac{\epsilon_0}{2}$, $R = \lceil \log_2 \frac{\epsilon_0}{\epsilon}\rceil$, and $T_1^{(1)} = 81Lc_0^2(1/\epsilon_0)^{1-2\theta}$. Let run SVRG$^{\text{HEB-RS}}$ with $S = \left\lceil \frac{1}{\frac{1}{2}-\theta}\log_2\left(\frac{c}{c_0}\right)\right\rceil + 1$, then $\mathrm{E}[F(x^{(S)}) - F_*] \le \epsilon$. The iteration complexity of SVRG$^{\text{HEB-RS}}$ is

$$O\left(n\log(\epsilon_0/\epsilon)\log(c/c_0) + \frac{Lc^2}{\epsilon^{1-2\theta}}\right).$$

Adaptive SVRG for $\theta = 1/2$

- **The challenge** is to decide when we should increase the value of c: In light of the value of T_1 in Theorem 2 for $\theta = 1/2$, i.e., $T_1 = \lceil 81Lc^2\rceil$, one might consider to start with a small value for c and then increase its value by a constant factor at certain points in order to increase the value of T_1.
- **The goal** is to develop an appropriate "certificate" that can be easily verified and can act as signal to check whether the value of c is already large enough for a sufficient decrease in the objective value.
- **The motivation** of the developed certificate is the property of proximal gradient update under the QEB, i.e.,
  $$F(\bar{x}) - F_* \le (L + L_f)^2c^2\|\bar{x} - \tilde{x}\|_2^2,$$
  where $\bar{x} = \arg\min_{x\in\Omega}\langle\nabla f(\tilde{x}), x - \tilde{x}\rangle + \frac{L}{2}\|x - \tilde{x}\|_2^2 + \Psi(x)$.
- The term $\|\bar{x} - \tilde{x}\|_2$ can be used as a gauge for monitoring the decrease in the objective value by performing the proximal gradient update. Although the full gradient is computationally expensive, SVRG allows to compute it at a small number of reference points.
- **Searching the value of c:** The full gradients are leveraged to develop the certificate for searching the value of c. The detailed steps are presented in Step 8 to Step 10 of Algorithm 3. If c_s is larger than c, the condition in Step 8 is true with small probability, which is stated in the following lemma.

Lemma 1. *Assume problem (1) satisfies the QEB condition. Let $\eta = \frac{1}{36L}$, $T_s = \lceil 81Lc_s^2\rceil$, $R_s = \left\lceil \log_2\left(\frac{2c_s^2(L+L_f)^2}{\vartheta^2\rho L}\right)\right\rceil$. Then for any $\vartheta \in (0, 1)$, we have*

$$\Pr\left(\|\bar{x}^{(s+1)} - \tilde{x}^{(s+1)}\|_2 \ge \vartheta\|\bar{x}^{(s)} - \tilde{x}^{(s)}\|_2 \,\middle|\, c_s \ge c\right) \le \rho.$$

Algorithm 3 SVRG under QEB with Restarting and Search: SVRG$^{\text{QEB-RS}}$

1: **Input**: $\tilde{x}^{(0)} \in \Omega$, an initial value $c_0 > 0$, $\epsilon > 0$, $\rho = \log^{-1}(1/\epsilon)$ and $\vartheta \in (0, 1)$.
2: $\bar{x}^{(0)} = \arg\min_{x\in\Omega}\langle\nabla f(\tilde{x}^0), x - \tilde{x}^0\rangle + \frac{L}{2}\|x - \tilde{x}^0\|_2^2 + \Psi(x)$, $s = 0$
3: **while** $\|\bar{x}^{(s)} - \tilde{x}^{(s)}\|_2^2 > \epsilon$ **do**
4: Set $T_s = \lceil 81Lc_s^2\rceil$ and $T_s = \log_{\frac{1}{9\delta}}\left(\frac{\kappa_s}{\vartheta\tau}\right)$
5: $\tilde{x}^{(s+1)}$=SVRG$^{\text{HEB}}$$(\bar{x}^{(s)}, T_s, R_s, 0.5)$
6: $\bar{x}^{(s+1)} = \arg\min_{x\in\Omega}\langle\nabla f(\tilde{x}^{(s+1)}), x - \tilde{x}^{(s+1)}\rangle + \frac{L}{2}\|x - \tilde{x}^{(s+1)}\|_2^2 + \Psi(x)$
7: $c_{s+1} = c_s$
8: **if** $\|\bar{x}^{(s+1)} - \tilde{x}^{(s+1)}\|_2 \ge \vartheta\|\bar{x}^{(s)} - \tilde{x}^{(s)}\|_2$ **then**
9:   $c_{s+1} = \sqrt{2}c_s$, $\bar{x}^{(s+1)} = \bar{x}^{(s)}$, $\tilde{x}^{(s+1)} = \tilde{x}^{(s)}$
10: **end if**
11: $s = s + 1$
12: **end while**
13: **Output**: $\bar{x}^{(s)}$

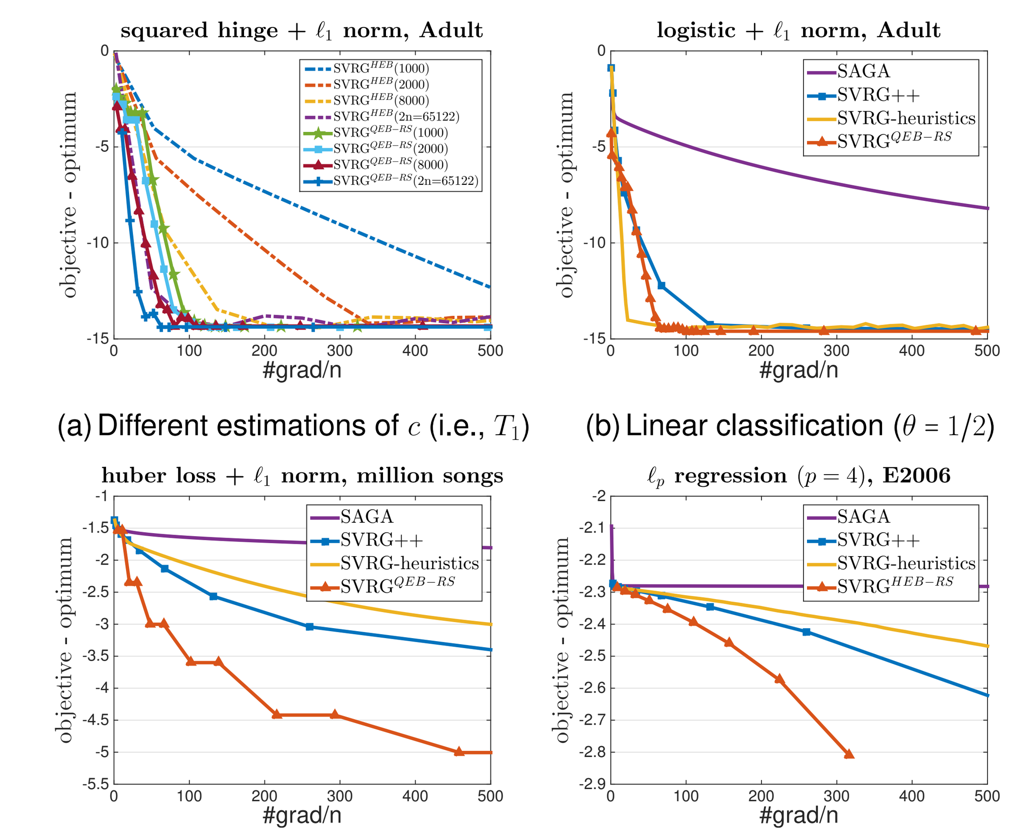
Main Result 2

Theorem 3. Assume problem (1) satisfies the QEB condition. Let $\rho = \log^{-1}(1/\epsilon)$, $\eta = \frac{1}{36L}$, $T_s = \lceil 81Lc_s^2\rceil$, and $R_s = \left\lceil \log_2\left(\frac{2c_s^2(L+L_f)^2}{\vartheta^2\rho L}\right)\right\rceil$. The expected iteration complexity of SVRG$^{\text{QEB-RS}}$ is

$$O\left((Lc^2 + n)\log_2\left(\frac{c^2(L+L_f)^2}{\vartheta^2 L}\log\left(\frac{1}{\epsilon}\right)\right)\left(\log_{1/\vartheta^2}\left(\frac{\|\bar{x}^{(0)} - \tilde{x}^{(0)}\|_2^2}{\epsilon}\right) + \log_2\left(\frac{c}{c_0}\right)\right)\right).$$

Applications and Experiments

1. Piecewise convex quadratic (PCQ) problems
   - Examples of loss function: square loss $\ell(z, b) = (z - b)^2$; squared hinge loss $\ell(z, b) = \max(0, 1 - bz)^2$; Huber loss $\ell_\gamma(z, b) = \begin{cases}\frac{1}{2}(z-b)^2 & \text{if } |z - b| \le \gamma,\\ \gamma(|z - b| - \frac{1}{2}\gamma) & \text{otherwise.}\end{cases}$
   - Examples of regularization: ℓ_1 norm, ℓ_∞ norm or $\ell_{1,\infty}$ norm regularization.
   - It satisfies the QEB condition, i.e., $\theta = 1/2$.
2. A family of structured smooth composite functions: $F(x) = h(Ax) + \Psi(x)$
   - $\Psi(x)$ is a polyhedral function or an indicator function of a polyhedral set.
   - $h(\cdot)$ is a smooth and strongly convex function on any compact set.
   - Examples of loss function: square loss $\ell(z, b) = (z - b)^2$; logistic loss $\ell(z, b) = \log(1 + \exp(-zb))$.
   - It satisfies the QEB condition, i.e., $\theta = 1/2$.
3. ℓ_1 constrained ℓ_p norm regression: $F(x) = 1/n\sum_{i=1}^n(x^\top a_i - b_i)^p$, where $p \in 2\mathbb{N}^+$.
   - It satisfies the HEB condition with intermediate values of $\theta \in (0, 1/2)$, i.e., $\theta = 1/p$.



(a) Different estimations of c (i.e., T_1)
(b) Linear classification ($\theta = 1/2$)
(c) Linear regression ($\theta = 1/2$)
(d) Linear regression ($\theta = 1/4$)

[1] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In NIPS, pages 315?323, 2013.
[2] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. SIAM Journal on Optimization, 24(4):2057?2075, 2014.
[3] Z. Allen-Zhu and Y. Yuan. Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In ICML, pages 1080?1089, 2016.
[4] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. Suter. From error bounds to the complexity of first-order descent methods for convex functions. CoRR, abs/1510.08234, 2015.
[5] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. arXiv:1602.06661, 2016.
[6] P. Gong and J. Ye. Linear convergence of variance-reduced projected stochastic gradient without strong convexity. CoRR, abs/1406.1102, 2014.
[7] H. Karimi, J. Nutini, and M. W. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In ECML-PKDD, pages 795?811, 2016.
[8] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non- strongly convex optimization. CoRR, abs/1504.06298, 2015.
[9] H. Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. CoRR, abs/1606.00269, 2016.