Towards closing the gap between the theory and practice of SVRG

Othmane Sebbouh, Nidham Gazagnadou, Samy Jelassi, Francis Bach, Robert M. Gower

h

Consider the optimization problem:

$$x^* = \arg\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f_i(x) =: f(x), \qquad (1)$$

where:

- f is L-smooth and μ-strongly convex,
- each f_i is $L_{\max}$-smooth.

Stochastic Variance Reduced Gradient

Algorithm 1 *SVRG [?]*

Parameters inner-loop length $m \gtrapprox \frac{L_{\max}}{\mu}$, step size α, $p_t := \frac{1}{m}$
Initialization $w_0 = x_0^m \in \mathbb{R}^d$
for $s = 1, 2, \dots$ **do**
  $x_s^0 = w_{s-1}$
  for $t = 0, 1, \dots, m-1$ **do**
    Sample i_t uniformly at random in $\{1, \dots, n\}$
    $g_s^t = \nabla f_{i_t}(x_s^t) - \nabla f_{i_t}(w_{s-1}) + \nabla f(w_{s-1})$
    $x_s^{t+1} = x_s^t - \alpha g_s^t$
  end for
  $w_s = \sum_{t=0}^{m-1} p_t x_s^t$
end for

Problem: SVRG [?] differs from practice on 3 important points:

- Constraint on the size of the loop m.
- First iterate reset to average of past iterates.
- No result showing benefits from mini-batching.

Motivations

- Develop an algorithm which is closer to practice.
- Offer strong theoretical guarantees on its convergence.
- Demonstrate benefits from mini-batching.

Stochastic Reformulation

Problem (1) can be reformulated as

$$x^* = \arg\min_{x\in\mathbb{R}^d} \mathbb{E}_{v\sim D}\left[\frac{1}{n}\sum_{i=1}^n v_i f_i(x)\right] := \mathbb{E}_{v\sim D}[f_v(x)], \qquad (2)$$

where $\mathbb{E}_{v\sim D}[v] = \mathbf{1}_n$. To solve (2), we can use SVRG:

$$x_s^{t+1} = x_s^t - \gamma\left(\nabla f_{v_t}(x_s^t) - \nabla f_{v_t}(w_{s-1}) + \nabla f(w_{s-1})\right),$$

where $v^k \sim \mathcal{D}$ is sampled at each iteration.
Arbitrary sampling allows to simultaneously analyze all possible forms of sampling.

Example: mini-batching without replacement

Consider a random set valued-map S which picks from all $\binom{n}{b}$ subsets of $\{1, \dots, n\}$ of size b. Let:

$$v_i = \begin{cases} \frac{n}{b} & \text{if } i \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Then: $f_v(x) = \frac{1}{b}\sum_{i\in S} f_i(x)$ and $\nabla f_v(x) = \frac{1}{b}\sum_{i\in S}\nabla f_i(x)$.

Proposed Algortihm: Free-SVRG

Algorithm 2 *Free-SVRG*

Parameters Free inner-loop length m, step size α, $p_t := \frac{(1-\alpha\mu)^{m-1-t}}{\sum_{i=0}^{m-1}(1-\alpha\mu)^{m-1-i}}$.
Initialization $w_0 = x_0^m \in \mathbb{R}^d$
for $s = 1, 2, \dots$ **do**
  $x_s^0 = x_{s-1}^m$
  for $t = 0, 1, \dots, m-1$ **do**
    Sample i_t uniformly at random in $\{1, \dots, n\}$
    $g_s^t = \nabla f_{i_t}(x_s^t) - \nabla f_{i_t}(w_{s-1}) + \nabla f(w_{s-1})$
    $x_s^{t+1} = x_s^t - \alpha g_s^t$
  end for
  $w_s = \sum_{t=0}^{m-1} p_t x_s^t$
end for

Solves several issues with SVRG [?]:

- Continuously updated iterates (no averaging).
- Free choice of the inner loop size.
- Much easier analysis.

Algorithm analysis

An essential constant for the analysis is the **expected smoothness**.

Lemma: Expected smoothness

Let $v \sim \mathcal{D}$ be a sampling vector. There exists $\mathcal{L} \geq 0$ such that for all $x \in \mathbb{R}^d$,

$$\mathbb{E}_{v\sim D}\left[\|\nabla f_v(x) - \nabla f_v(x^*)\|_2^2\right] \leq 2\mathcal{L}\left(f(x) - f(x^*)\right).$$

Example: for **mini-batching without replacement**,

$$\mathcal{L} = \mathcal{L}(b) = \frac{1}{b}\frac{n-b}{n-1}L_{\max} + \frac{n}{b}\frac{b-1}{n-1}L.$$

In particular: $\mathcal{L}(1) = L_{\max}$ and $\mathcal{L}(n) = L$.

Convergence Theorem 1

Consider the setting of Algorithm 2 and the following Lyapunov function

$$\phi_s := \|x_s^m - x^*\|_2^2 + 8\alpha^2\mathcal{L}S_m(f(w_s) - f(x^*)).$$

If $\alpha \leq \frac{1}{6\mathcal{L}}$, then

$$\mathbb{E}[\phi_s] \leq \beta^s\phi_0, \quad \text{where} \quad \beta = \max\left\{(1-\alpha\mu)^m, \tfrac{1}{2}\right\}.$$

Total complexity

The **total complexity** of finding an $\epsilon > 0$ approximate solution that satisfies $\mathbb{E}\left[\|x_s^m - x^*\|_2^2\right] \leq \epsilon\,\phi_0$ is

$$C_m(b) := 2\left(\frac{n}{m} + 2b\right)\max\left\{\frac{3\mathcal{L}(b)}{\mu}, m\right\}\log\left(\frac{1}{\epsilon}\right)$$

And for **mini-batching** (dropping the log term):

$$C_m(b) := 2\left(\frac{n}{m} + 2b\right)\max\left\{\frac{3}{b}\frac{n-b}{n-1}\frac{L_{\max}}{\mu} + \frac{3n}{b}\frac{b-1}{n-1}\frac{L}{\mu}, m\right\}.$$

How to set the inner loop size?

Since m is not constrained, we can choose the one that minimizes the total complexity.
Answer: There is a **range of values** that minimize the complexity.

$$m \in \left[\min(n, \tfrac{L_{\max}}{\mu}), \max(n, \tfrac{L_{\max}}{\mu})\right] \implies O\left(\left(n + \frac{L_{\max}}{\mu}\right)\log\frac{1}{\epsilon}\right)$$

Rem: Includes the practical choice n !

Alternative algorithm: L-SVRG-D

Problem: SVRG relies on knowing μ.
Solution: [?] proposed a loopless version of SVRG.
Improvement: Decrease the step size when the variance of the gradient is high.

Algorithm 3 L-SVRG-D

Parameters step size α, $p \in (0, 1]$.
Initialization $w^0 = x^0 \in \mathbb{R}^d$, $\alpha_0 = \alpha$
for $k = 0, 1, 2, \dots$ **do**
  Sample $v_k \sim \mathcal{D}$
  $g^k = \nabla f_{v_k}(x^k) - \nabla f_{v_k}(w^k) + \nabla f(w^k)$
  $x^{k+1} = x^k - \alpha_k g^k$
  $(w^{k+1}, \alpha_{k+1}) = \begin{cases} (x^k, \alpha) & \text{with probability } p \\ (w^k, \sqrt{1-p}\,\alpha_k) & \text{with probability } 1-p \end{cases}$
end for

Convergence Theorem 2

Consider the iterates of Algorithm 3 and the following Lyapunov function

$$\phi^k := \|x^k - x^*\|_2^2 + \frac{8\alpha_k^2\mathcal{L}}{p(3-2p)}\left(f(w^k) - f(x^*)\right).$$

If $p \approx \frac{1}{n}$ and $\alpha \lessapprox \frac{2}{7\mathcal{L}}$, then

$$\mathbb{E}\left[\phi^k\right] \leq \beta^k\phi^0, \quad \text{where} \quad \beta = \max\left\{1 - \frac{2}{3}\alpha\mu, 1 - \frac{p}{2}\right\}.$$

Benefits:

- Bigger step size in the beginning of the loop, when the variance is low.
- Smaller step size in the end of loop, when the variance is high.

Total complexity, optimal inner loop size: similar to those of *Free-SVRG* up to constants.

Optimal mini-batch size

We determine the optimal mini-batch size for *Free-SVRG* and *L-SVRG-D* for the usual choice $m = n$ (or $p = \frac{1}{n}$):

$$b^* = \begin{cases} 1 & \text{if } n \geq \frac{3L_{\max}}{\mu} \\ \left\lfloor \min(\tilde{b}, \hat{b}) \right\rfloor & \text{if } \frac{3L}{\mu} < n < \frac{3L_{\max}}{\mu} \\ \left\lfloor \hat{b} \right\rfloor & \text{otherwise, if } n \leq \frac{3L}{\mu} \end{cases}$$

$$\text{where} \quad \begin{aligned} \hat{b} &:= \sqrt{\frac{n}{2}\frac{L_{\max} - L}{nL - L_{\max}}} \\ \tilde{b} &:= \frac{3n(L_{\max} - L)}{n(n-1)\mu - 3(nL - L_{\max})} \end{aligned}$$

References