Nonconvex Variance Reduced Optimization with Arbitrary Sampling

Samuel Horváth 1 Peter Richtárik 1, 2, 3
1 KAUST, 2University of Edinburgh, 3Moscow Institute of Physics and Technology

The Problem

\[ \min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) \]  
(1)

- \( f_i \) is \( L_i \)-smooth but non-convex
- \( n \) is big

Arbitrary Sampling

- Sampling: a random set-valued mapping \( S \) with values being subsets of \([n] := \{1, 2, \ldots, n\}\). A sampling is used to generate minibatches in each iteration.
- Probability matrix associated with sampling \( S \) :
  \[ P_{ij} \overset{\text{def}}{=} \text{Prob}(i \subseteq S) \]
- Probability vector associated with sampling \( S \) :
  \[ P = (\mathbb{P}_1, \ldots, \mathbb{P}_n) \]
- Minibatch size: \( b = \mathbb{E}[|S|] \) (expected size of \( S \))
- Proper sampling: Sampling for which \( \mathbb{P}_i > 0 \) for all \( i \in [n] \)
- “Arbitrary sampling” = any proper sampling

Key Lemma

Let \( \zeta_1, \zeta_2, \ldots, \zeta_n \) be vectors in \( \mathbb{R}^d \) and let \( \tilde{\zeta} \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \zeta_i \) be their average. Let \( S \) be a proper sampling. Let \( v = (v_1, \ldots, v_n) \) > 0 be such that

\[ P - pp^T \preceq \text{Diag}(v_1, \ldots, v_n) \]

Then

\[ \mathbb{E} \left[ \frac{1}{b} \sum_{S' \in S} \| \frac{1}{b} \sum_{i \in S'} \zeta_i - \frac{1}{n} \sum_{i=1}^n \zeta_i \|^2 \right] \leq \frac{n}{b n^2} \sum_{i=1}^n v_i \| \zeta_i \|^2. \]

Whenever (2) holds, it must be the case that

\[ v_i \geq 1 - \mathbb{P}_i, \]

Optimal Sampling & Superlinear Speedup

- Under our analysis, the independent sampling \( S^* \) defined by
  \[ \mathbb{P}_i \overset{\text{def}}{=} \begin{cases} (b + k - n) \frac{L_i}{\sum_{j=1}^n L_j} & \text{if } i \leq k \\ 1 & \text{if } i > k \end{cases} \]
  is optimal, where \( k \) is the largest integer satisfying \( 0 < b + k - n \leq \frac{L_i}{\sum_{j=1}^n L_j} \).
- All 3 methods enjoy superlinear speed in \( b \) up to the minibatch size \( b_{\text{max}} := \max \{b \mid bL_i \leq \sum_{j=1}^n L_j \} \).

Main Contributions

- We develop arbitrary sampling variants of 3 popular variance-reduced methods for solving the non-convex problem (1): SVRG [1], SAGA [2], SARAH [3].
- Our rates for \( b = 1 \): up to \( n \times \text{better} \) (depending on \{\( L_i \}\}). Improvements even in the case when \( L_i = L_1 \) for all \( i, j \) (for SVRG & SAGA).
- Our rates for \( b \geq 1 \): Linear or superlinear speedup in minibatch size \( b \). That is, # of iterations needed to output a solution of a given accuracy drops by a factor equal or greater to \( b \).
- We design importance sampling & approximate importance sampling for minibatches, which vastly outperform standard uniform minibatch strategies in practice.

Numerical Results

SRVRG with Arbitrary Sampling

Algorithm 1: SVRG \((x^0, m, T, \eta, S)\)

\[ x^0 = x^0_m = x^m, \quad M = \frac{T}{m}; \]

for \( s = 0 \) to \( M - 1 \) do

\[ x^{s+1} = x^m + g^{s+1} - \frac{1}{m} \sum_{i=1}^m \nabla f_i(x^m) \]

for \( t = 0 \) to \( m - 1 \) do

\[ \{y^t_i\} = \text{Random subset (minibatch) of cardinality } S \]

\[ y^t_i = \sum_{S \in S} \left( \nabla f_i(x^{s+1}) - \nabla f_i(y^t_i) \right) + g^{s+1} \]

\[ x^{s+1} = x^{s+1} + \eta y^t_i \]

end

end

Output: \( x_m \) chosen uniformly random from \( \{x^{s+1}_i\}_{i=0}^M \)

# Stochastic Gradient Evaluations to Achieve \( \mathbb{E} \| \nabla f(x) \| \leq \epsilon \)

<table>
<thead>
<tr>
<th>Alg</th>
<th>Uniform sampling</th>
<th>Arbitrary sampling [NEW]</th>
<th>SVRG [NEW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVRG</td>
<td>( n + \frac{4(\log m)^2}{\epsilon^2} ) [1]</td>
<td>( n + \frac{(\log m)^2}{\epsilon^2} )</td>
<td>( n + \frac{(\log m)^2}{\epsilon^2} )</td>
</tr>
<tr>
<td>SAGA</td>
<td>( n + \frac{2(\log m)^2}{\epsilon} ) [2]</td>
<td>( n + \frac{(\log m)^2}{\epsilon} )</td>
<td>( n + \frac{(\log m)^2}{\epsilon} )</td>
</tr>
<tr>
<td>SARAH</td>
<td>( n + \frac{4(\log m)^2}{\epsilon} ) [3]</td>
<td>( n + \frac{(\log m)^2}{\epsilon} )</td>
<td>( n + \frac{(\log m)^2}{\epsilon} )</td>
</tr>
</tbody>
</table>

Constants: \( L_{\text{max}} = \max L_i, L = \frac{1}{n} \sum L_i \), \( \alpha \overset{\text{def}}{=} \frac{1}{2} n \sum_{i=1}^n \frac{L_i}{\sum_{j=1}^n L_j} \)

References

