Nonconvex Variance Reduced Optimization with Arbitrary Sampling

Samuel Horváth[1] Peter Richtárik[1,2,3]

[1] KAUST [2]University of Edinburgh [3]Moscow Institute of Physics and Technology

The Problem

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

- 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 :
$$\mathbf{P}_{ij} \overset{\text{def}}{=} \text{Prob}(\{i, j\} \subseteq S)$$
- Probability vector associated with sampling S :
$$p = (p_1, \ldots, p_n), \quad p_i \overset{\text{def}}{=} \text{Prob}(i \in S)$$
- Minibatch size: $b = \mathrm{E}[|S|]$ (expected size of S)
- Proper sampling: Sampling for which $p_i > 0$ for all $i \in [n]$
- "Arbitrary sampling" = any proper sampling

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].
- We are able calculate the optimal sampling out of all samplings of a given minibatch size. This is the first time an optimal minibatch sampling was computed (from the class of all samplings).
- We design importance sampling & approximate importance sampling for minibatches, which vastly outperform standard uniform minibatch strategies in practice.

Key Lemma

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

$$\mathbf{P} - pp^\top \preceq \mathbf{Diag}(p_1 v_1, p_2 v_2, \ldots, p_n v_n). \qquad (2)$$

Then

$$\mathrm{E}\left[\left\|\sum_{i\in S}\frac{\zeta_i}{np_i} - \bar\zeta\right\|^2\right] \le \frac{1}{n^2}\sum_{i=1}^n \frac{v_i}{p_i}\|\zeta_i\|^2.$$

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

$$v_i \ge 1 - p_i.$$

Optimal Sampling & Superlinear Speedup

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

Stochastic Gradient Evaluations to Achieve $\mathrm{E}\left[\|\nabla f(x)\|^2\right] \le \epsilon$

Alg	Uniform sampling	Arbitrary sampling [NEW]	S^* (Best Sampling) [NEW]
SVRG	$\max\left\{n, \frac{(1+4/3)L_{\max}c_1 n^{2/3}}{\epsilon}\right\}$ [1]	$\max\left\{n, \frac{(1+4\alpha/3)\bar L c_1 n^{2/3}}{\epsilon}\right\}$	$\max\left\{n, \frac{\left(1+\frac{4(n-b)}{3n}\right)\bar L c_1 n^{2/3}}{\epsilon}\right\}$
SAGA	$n + \frac{2L_{\max}c_2 n^{2/3}}{\epsilon}$ [2]	$n + \frac{(1+\alpha)\bar L c_2 n^{2/3}}{\epsilon}$	$n + \frac{(1+\frac{n-b}{n})\bar L c_2 n^{2/3}}{\epsilon}$
SARAH	$n + \frac{\frac{n-b}{n-1}L_{\max}^2 c_3}{\epsilon^2}$ [3]	$n + \frac{\alpha \bar L^2 c_3}{\epsilon^2}$	$n + \frac{\frac{n-b}{n}\bar L^2 c_3}{\epsilon^2}$

Constants: $L_{\max} = \max_i L_i$ $\quad \bar L = \frac{1}{n}\sum_i L_i$ $\quad c_1, c_2, c_3 =$ universal constants $\quad \alpha := \frac{b}{\bar L^2 n^2}\sum_{i=1}^n \frac{v_i L_i^2}{p_i}$

Samplings

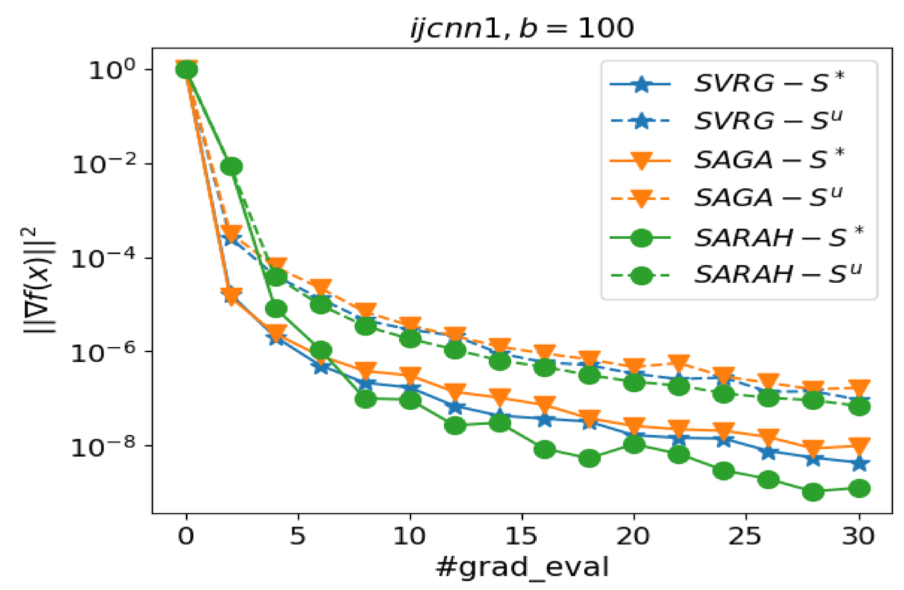
- **Uniform** S^u: Every subset of $[n]$ of size b (minibatch size) is chosen with the same probability: $1/\binom{n}{b}$
- **Independent** S^*: For each $i \in [n]$ we independently flip a coin, and with probability p_i include element i into S.
- **Approximate Independent** S^a: Fix some $k \in [n]$ and let $a = \lceil k \max_{i\le k} p_i \rceil$. We now sample a single set S' of cardinality a using the uniform minibatch sampling S^u. Subsequently, we apply an independent sampling S^* to select elements of S', with selection probabilities $p'_i = kp_i/a$. The resulting random set is S^a.

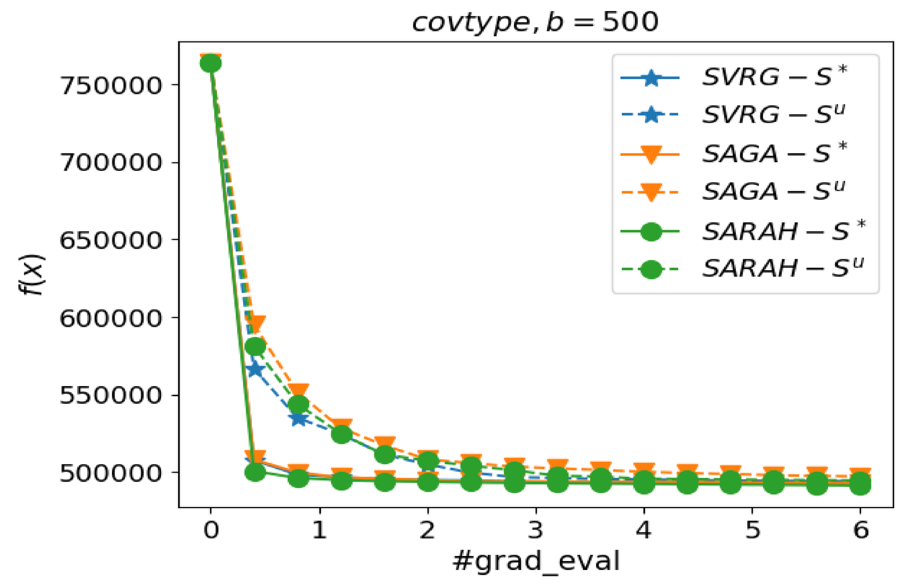
SVRG with Arbitrary Sampling

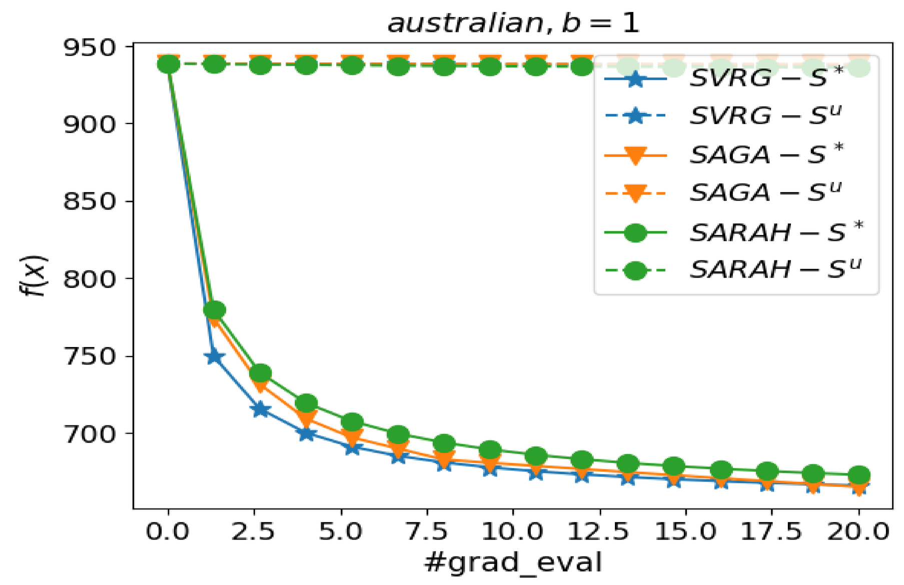
Algorithm 1: SVRG

$\tilde x^0 = x_m^0 = x^0$, $M = \lceil T/m \rceil$;
for $s = 0$ *to* $M - 1$ **do**
  $x_0^{s+1} = x_m^s$; $g^{s+1} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde x^s)$
  for $t = 0$ *to* $m - 1$ **do**
    Draw a random subset (minibatch) $S_t \sim S$
    $v_t^{s+1} = \sum_{i_t\in S_t}\frac{1}{np_{i_t}}(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde x^s)) + g^{s+1}$
    $x_{t+1}^{s+1} = x_t^{s+1} - \eta v_t^{s+1}$
  end
  $\tilde x^{s+1} = x_m^{s+1}$
end
Output: Iterate x_a chosen uniformly random from $\{\{x_t^{s+1}\}_{t=0}^m\}_{s=0}^M$

Numerical Results







References

[1] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *The 33th International Conference on Machine Learning*, pages 314–323, 2016.

[2] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Fast incremental method for smooth nonconvex optimization. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 1971–1977. IEEE, 2016.

[3] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv:1705.07261*, 2017.