ICML | 2019

Thirty-sixth International Conference on Machine Learning

Nonconvex Variance Reduced Optimization with Arbitrary Sampling

Samuel Horváth

KAUST
King Abdullah University of Science and Technology

Peter Richtárik

KAUST
King Abdullah University of Science and Technology

The University of Edinburgh

MIPT
Moscow Institute of Physics and Technology

Empirical Risk Minimization

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

Empirical Risk Minimization

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

n is big

Empirical Risk Minimization

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

n is big

non-convex, L_i-smooth

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|$$

Baseline Variance Reduced SGD Methods

SVRG

Johnson & Zhang
NIPS 2013

$$x^+ = x - \eta\left(\nabla f_i(x) - \nabla f_i(\hat{x}) + \nabla f(\hat{x})\right)$$

SAGA

Defazio, Bach & Lacoste-Julien
NIPS 2014

$$x^+ = x - \eta\left(\nabla f_i(x) - g_i + \frac{1}{n}\sum_{j=1}^{n} g_j\right)$$

SARAH

Nguyen, Liu, Scheinberg & Takáč
ICML 2017

$$x^+ = x - \eta\left(\nabla f_i(x) - \nabla f_i(x^-) + v^-\right)$$

Baseline Variance Reduced SGD Methods

SVRG

Johnson & Zhang
NIPS 2013

$$x^+ = x - \eta\left(\nabla f_i(x) - \nabla f_i(\hat{x}) + \nabla f(\hat{x})\right)$$

Uniform sampling

SAGA

Defazio, Bach & Lacoste-Julien
NIPS 2014

$$x^+ = x - \eta\left(\nabla f_i(x) - g_i + \frac{1}{n}\sum_{j=1}^{n} g_j\right)$$

Uniform sampling

SARAH

Nguyen, Liu, Scheinberg & Takáč
ICML 2017

$$x^+ = x - \eta\left(\nabla f_i(x) - \nabla f_i(x^-) + v^-\right)$$

Uniform sampling

Baseline Variance Reduced SGD Methods–Mini-batch

SVRG

Konečný & Richtárik
FAMS 2017

$$x^+ = x - \eta \left(\frac{1}{b} \sum_{i \in S} (\nabla f_i(x) - \nabla f_i(\hat{x})) + \nabla f(\hat{x}) \right)$$

SAGA

Reddi, Hefny, Sra, Poczos, Smola
CDC 2016

$$x^+ = x - \eta \left(\frac{1}{b} \sum_{i \in S} (\nabla f_i(x) - g_i) + \frac{1}{n} \sum_{j=1}^{n} g_j \right)$$

SARAH

Nguyen, Liu, Scheinberg & Takáč
2017

$$x^+ = x - \eta \left(\frac{1}{b} \sum_{i \in S} (\nabla f_i(x) - \nabla f_i(x^-)) + v^- \right)$$

Baseline Variance Reduced SGD Methods–Mini-batch

Mini-batch size

SVRG

Konečný & Richtárik
FAMS 2017

$$x^+ = x - \eta \left(\frac{1}{b} \sum_{i \in S} (\nabla f_i(x) - \nabla f_i(\hat{x})) + \nabla f(\hat{x}) \right)$$

Uniform sampling

SAGA

Reddi, Hefny, Sra, Poczos, Smola
CDC 2016

$$x^+ = x - \eta \left(\frac{1}{b} \sum_{i \in S} (\nabla f_i(x) - g_i) + \frac{1}{n} \sum_{j=1}^{n} g_j \right)$$

Uniform sampling

SARAH

Nguyen, Liu, Scheinberg & Takáč
2017

$$x^+ = x - \eta \left(\frac{1}{b} \sum_{i \in S} (\nabla f_i(x) - \nabla f_i(x^-)) + v^- \right)$$

Uniform sampling

Contributions

- Analysis of SVRG, SAGA and SARAH in the arbitrary sampling paradigm
- Construction of optimal minibatch sampling

Contributions

Richtárik & Takáč (OL 2016; arXiv 2013)
Qu, Richtárik & Zhang (NIPS 2015)
Qu & Richtárik (COAP 2016)
Chambolle, Ehrhardt, Richtárik & Schoenlieb (SIOPT 2018)
Hanzely & Richtárik (AISTATS 2019)
Qian, Qu & Richtárik (ICML 2019)
Gower, Loizou, Qian, Sailanbayev, Shulgin & Richtárik (ICML 2019)

- Analysis of SVRG, SAGA and SARAH in the arbitrary sampling paradigm

- Construction of optimal minibatch sampling

Contributions

Richtárik & Takáč (OL 2016; arXiv 2013)
Qu, Richtárik & Zhang (NIPS 2015)
Qu & Richtárik (COAP 2016)
Chambolle, Ehrhardt, Richtárik & Schoenlieb (SIOPT 2018)
Hanzely & Richtárik (AISTATS 2019)
Qian, Qu & Richtárik (ICML 2019)
Gower, Loizou, Qian, Sailanbayev, Shulgin & Richtárik (ICML 2019)

- Analysis of SVRG, SAGA and SARAH in the arbitrary sampling paradigm

- Construction of optimal minibatch sampling

First optimal/importance sampling for minibatches!

Data Sampling (i.e., Mini-batching) Mechanisms

Sampling: a random subset of {1, 2, … , *n*}

Probability matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ associated with sampling S

$$\mathbf{P}_{ij} := \mathrm{Prob}(\{i, j\} \subseteq S)$$

Probability vector $p \in \mathbb{R}^n$ associated with sampling S

$$p_i := \mathrm{Prob}(\{i\} \subseteq S) = \mathbf{P}_{ii}$$

Proper sampling: $p_i > 0$ for all $i = 1, 2, \ldots, n$

Data Sampling (i.e., Mini-batching) Mechanisms

A **sampling** is uniquely defined by assigning probabilities to all 2^n subsets of {1, 2, ... , *n*}

Sampling: a random subset of {1, 2, ... , *n*}

Probability matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ associated with sampling S

$$\mathbf{P}_{ij} := \text{Prob}(\{i,j\} \subseteq S)$$

Probability vector $p \in \mathbb{R}^n$ associated with sampling S

$$p_i := \text{Prob}(\{i\} \subseteq S) = \mathbf{P}_{ii}$$

Proper sampling: $p_i > 0$ for all $i = 1, 2, \ldots, n$

Data Sampling (i.e., Mini-batching) Mechanisms

A **sampling** is uniquely defined by assigning probabilities to all 2^n subsets of {1, 2, ... , *n*}

Sampling: a random subset of {1, 2, ... , *n*}

Probability matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ associated with sampling S

$$\mathbf{P}_{ij} := \mathrm{Prob}(\{i,j\} \subseteq S)$$

Probability vector $p \in \mathbb{R}^n$ associated with sampling S

$$p_i := \mathrm{Prob}(\{i\} \subseteq S) = \mathbf{P}_{ii}$$

Proper sampling: $p_i > 0$ for all $i = 1, 2, \ldots, n$

Examples

Standard sampling:

$S = \{i\}$ with probability $\frac{1}{n}$ for all $i = 1, 2, \ldots, n$

Standard mini-batch sampling:

$S = C$ with probability $\frac{1}{\binom{n}{b}}$ for all $C \subset \{1, 2, \ldots, n\}$ such that $|C| = b$

Data Sampling (i.e., Mini-batching) Mechanisms

A **sampling** is uniquely defined by assigning probabilities to all 2^n subsets of {1, 2, … , *n*}

Sampling: a random subset of {1, 2, … , *n*}

Probability matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ associated with sampling S

$$\mathbf{P}_{ij} := \mathrm{Prob}(\{i, j\} \subseteq S)$$

Probability vector $p \in \mathbb{R}^n$ associated with sampling S

$$p_i := \mathrm{Prob}(\{i\} \subseteq S) = \mathbf{P}_{ii}$$

Proper sampling: $p_i > 0$ for all $i = 1, 2, \ldots, n$

Examples

Standard sampling:

$S = \{i\}$ with probability $\frac{1}{n}$ for all $i = 1, 2, \ldots, n$

Standard mini-batch sampling:

$S = C$ with probability $\frac{1}{\binom{n}{b}}$ for all $C \subset \{1, 2, \ldots, n\}$ such that $|C| = b$

Arbitrary sampling paradigm = perform iteration complexity analysis for **any proper sampling**

From Standard Sampling to Arbitrary Sampling

SVRG with
Arbitrary Sampling

$$x^+ = x - \eta \left(\sum_{i \in S} \frac{1}{np_i} (\nabla f_i(x) - \nabla f_i(\hat{x})) + \nabla f(\hat{x}) \right)$$

From Standard Sampling to Arbitrary Sampling

SVRG with
Arbitrary Sampling

Unbiased estimator of
the gradient

$$x^+ = x - \eta \left(\overbrace{\sum_{i \in S} \frac{1}{np_i} \left(\nabla f_i(x) - \nabla f_i(\hat{x}) \right) + \nabla f(\hat{x})} \right)$$

From Standard Sampling to Arbitrary Sampling

SVRG with
Arbitrary Sampling

Unbiased estimator of
the gradient

$$x^+ = x - \eta \left(\sum_{i \in S} \frac{1}{n p_i} (\nabla f_i(x) - \nabla f_i(\hat{x})) + \nabla f(\hat{x}) \right)$$

Arbitrary
sampling

Standard sampling:

$$p_i = \frac{1}{n} \text{ for all } i$$

Convergence Rate I

$$\mathcal{O}\left(n + \alpha\frac{n^{2/3}}{\epsilon}\right)$$

Convergence Rate I

data points

$$\mathcal{O}\left(n + \alpha \frac{n^{2/3}}{\epsilon}\right)$$

$\mathrm{E}\|\nabla f(x)\|^2 \leq \epsilon$

Convergence Rate I

data points

$$\mathcal{O}\left(n + \alpha \frac{n^{2/3}}{\epsilon}\right)$$

$\mathrm{E}\|\nabla f(x)\|^2 \leq \epsilon$

KEY QUANTITY: DEPENDS ON THE SAMPLING

Convergence Rate II

$$\alpha := \frac{b}{\bar{L}n^2} \sum_{i=1}^{n} \frac{v_i L_i^2}{p_i}$$

Convergence Rate II

Constants satisfying:

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

$$\mathbf{P}_{ij} := \mathrm{Prob}(\{i,j\} \subseteq S)$$

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|$$

Expected mini-batch size:

$$b = \mathrm{E}|S|$$

$$\alpha := \frac{b}{\bar{L} n^2} \sum_{i=1}^{n} \frac{v_i L_i^2}{p_i}$$

$$\bar{L} = \frac{1}{n} \sum_{i=1}^{n} L_i$$

data points

$$p_i := \mathrm{Prob}(\{i\} \subseteq S) = \mathbf{P}_{ii}$$

Convergence Rate II

Constants satisfying:

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

$$\mathbf{P}_{ij} := \text{Prob}(\{i,j\} \subseteq S)$$

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|$$

Expected mini-batch size:

$$b = \text{E}|S|$$

$$\alpha := \frac{b}{\bar{L} n^2} \sum_{i=1}^n \frac{v_i L_i^2}{p_i}$$

$$\bar{L} = \frac{1}{n} \sum_{i=1}^n L_i$$

data points

$$p_i := \text{Prob}(\{i\} \subseteq S) = \mathbf{P}_{ii}$$

Optimal rate: minimize α over $\{(v_i, p_i)\}_{i=1}^n$

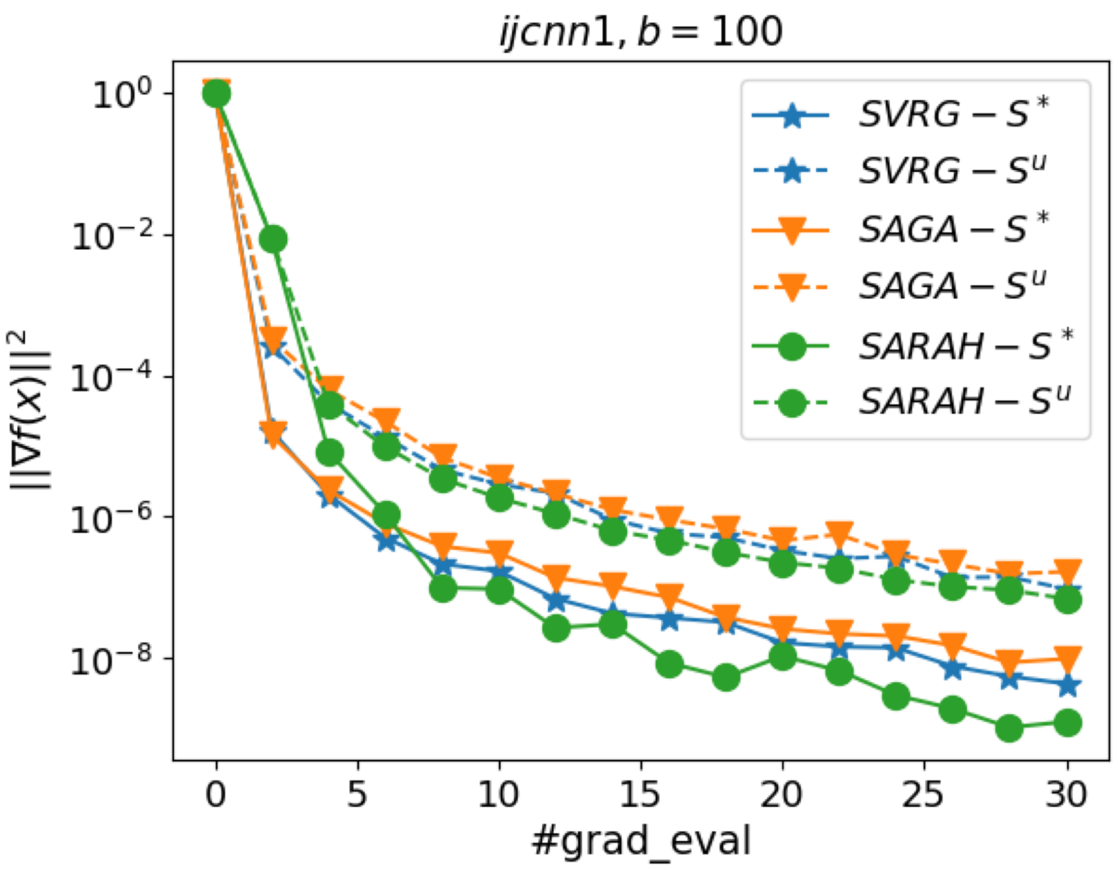
Convergence Rate III

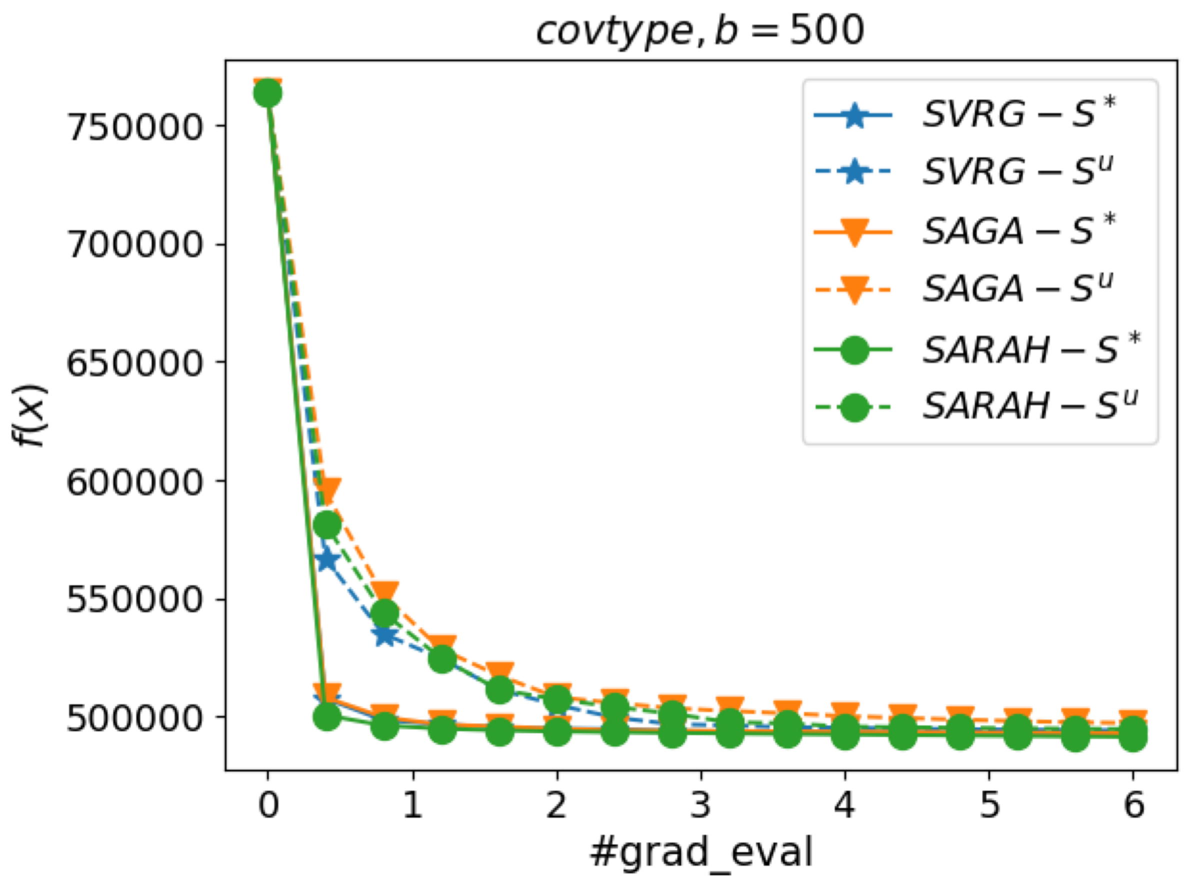
Stochastic Gradient Evaluations to Achieve $\mathrm{E}[\|\nabla f(x)\|^2] \leq \epsilon$

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 $\quad \alpha := \frac{b}{\bar{L}^2 n^2}\sum_{i=1}^n \frac{v_i L_i^2}{p_i}$

Experiments





Nonconvex Variance Reduced Optimization with Arbitrary Sampling

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

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

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

THE UNIVERSITY OF EDINBURGH

MIPT 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 $[n] := \{1, 2, \dots, n\}$. A sampling is used to generate minibatches in each iteration.
- Probability matrix associated with sampling S :
$$\mathbf{P}_{ij} \stackrel{\text{def}}{=} \text{Prob}(\{i,j\} \subseteq S)$$
- Probability vector associated with sampling S :
$$p = (p_1, \dots, p_n), \quad p_i \stackrel{\text{def}}{=} \text{Prob}(i \in S)$$
- 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 $\zeta_1, \zeta_2, \dots, \zeta_n$ be vectors in $\mathbb{R}^d$ and let $\bar{\zeta} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \zeta_i$ be their average. Let S be a proper sampling. Let $v = (v_1, \dots, v_n) > 0$ be such that

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

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
$$p_i \stackrel{\text{def}}{=} \begin{cases} (b+k-n)\frac{L_i}{\sum_{j=1}^k L_j}, & \text{if } i \le k \\ 1, & \text{if } i > k \end{cases},$$
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]	$n + \frac{(1+\alpha)\bar{L}c_2 n^{2/3}}{\epsilon}$	$n + \frac{\left(1+\frac{n-b}{n}\right)\bar{L}c_2 n^{2/3}}{\epsilon}$
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

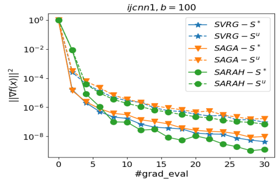
- **Uniform** S^u: Every subset of $[n]$ of size b (minibatch size) is chosen with the same probability: $1/\binom{n}{b}$
- **Independent** S^*: For each $i \in [n]$ we independently flip a coin, and with probability p_i include element i into S.
- **Approximate Independent** S^a: Fix some $k \in [n]$ 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.

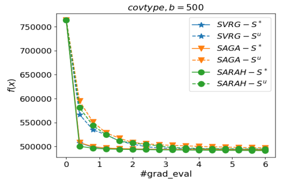
SVRG with Arbitrary Sampling

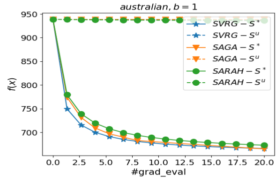
Algorithm 1: SVRG

$\tilde{x}^0 = x_m^0 = x^0$, $M = \lceil T/m \rceil$;
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) $S_t \sim S$
  $v_t^{s+1} = \sum_{i_t\in S_t}\frac{1}{np_{i_t}}\left(\nabla f_{i_t}(x_t^{s+1}) - \nabla f_{i_t}(\tilde{x}^s)\right) + g^{s+1}$
  $x_{t+1}^{s+1} = x_t^{s+1} - \eta v_t^{s+1}$
 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.

Poster: Pacific Ballroom #95 (Today 6:30–9:00 PM)

Thank you!