SAGA with Arbitrary Sampling

Xun Qian
Zheng Qu
Peter Richtárik
The Problem
The Problem: Regularized Empirical Risk Minimization

\[
\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)
\]
The Problem: Regularized Empirical Risk Minimization

\[
\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)
\]
The Problem: Regularized Empirical Risk Minimization

\[
\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)
\]
The Problem: Regularized Empirical Risk Minimization

\[
\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)
\]

- \( f(x) \): Training data
- \( \lambda_i \): Regularization parameters
- \( \psi(x) \): Regularizer
- \# training data

Regularizer
The Problem: Regularized Empirical Risk Minimization

\[
\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)
\]

- \( \lambda_i \): Weight associated with data point \( i \)
- \( f_i(x) \): Loss associated with data point \( i \)
- \( \psi(x) \): Regularizer
- \# training data
The Problem: Regularized Empirical Risk Minimization

$$\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)$$

- # training data
- Weight associated with data point $i$
- Loss associated with data point $i$
- Parameters describing the model
- Regularizer

$f(x)$
Arbitrary Sampling
SGD with Arbitrary Sampling
SGD with Arbitrary Sampling

1. In iteration $k$, we have $x^k$ available
In iteration $k$, we have $x^k$ available

Sample a random set $S_k \subseteq \{1, 2, \ldots, n\}$
SGD with Arbitrary Sampling

1. In iteration $k$, we have $x^k$ available
2. Sample a random set $S_k \subseteq \{1, 2, \ldots, n\}$
3. Compute the gradients $\nabla f_i(x^k)$ for $i \in S_k$ only
SGD with Arbitrary Sampling

1. In iteration $k$, we have $x^k$ available
2. Sample a random set $S_k \subseteq \{1, 2, \ldots, n\}$
3. Compute the gradients $\nabla f_i(x^k)$ for $i \in S_k$ only
4. Approximate the gradient $\nabla f(x^k)$ using $\{\nabla f_i(x^k) : i \in S_k\}$
SGD with Arbitrary Sampling

1. In iteration $k$, we have $x^k$ available
2. Sample a random set $S_k \subseteq \{1, 2, \ldots, n\}$
3. Compute the gradients $\nabla f_i(x^k)$ for $i \in S_k$ only
4. Approximate the gradient $\nabla f(x^k)$ using $\{\nabla f_i(x^k) : i \in S_k\}$
5. Take a stochastic gradient descent step to obtain $x^{k+1}$
### SGD with Arbitrary Sampling

1. In iteration $k$, we have $x^k$ available
2. Sample a random set $S_k \subseteq \{1, 2, \ldots, n\}$
3. Compute the gradients $\nabla f_i(x^k)$ for $i \in S_k$ only
4. Approximate the gradient $\nabla f(x^k)$ using $\{\nabla f_i(x^k) : i \in S_k\}$
5. Take a stochastic gradient descent step to obtain $x^{k+1}$

**Arbitrary sampling paradigm** (R. & Takáč 2013): want to be able to sample from any distribution over all $2^n$ subsets of $\{1, 2, \ldots, n\}$

- $p_i \overset{\text{def}}{=} \text{Prob}(i \in S_k)$
- $p_i > 0$ for all $i = 1, 2, \ldots, n$
Arbitrary Sampling: Examples for $n = 3$

$S_k = \{1, 2, 3\}$ with prob 1
Arbitrary Sampling: Examples for $n = 3$

**GD**

$S_k = \{1, 2, 3\}$ with prob 1

**SAGA**

$S_k = \{1\}$ with prob $1/3$

$S_k = \{2\}$ with prob $1/3$

$S_k = \{3\}$ with prob $1/3$
Arbitrary Sampling: Examples for $n = 3$

GD

\[ S_k = \{1, 2, 3\} \text{ with prob } 1 \]

SAGA

\[ S_k = \{1\} \text{ with prob } \frac{1}{3} \]
\[ S_k = \{2\} \text{ with prob } \frac{1}{3} \]
\[ S_k = \{3\} \text{ with prob } \frac{1}{3} \]

SAGA with nonuniform sampling

\[ S_k = \{1\} \text{ with prob } p_1 \]
\[ S_k = \{2\} \text{ with prob } p_2 \]
\[ S_k = \{3\} \text{ with prob } p_3 \]
Arbitrary Sampling: Examples for $n = 3$

**GD**

$S_k = \{1, 2, 3\}$ with prob 1

**SAGA**

$S_k = \{1\}$ with prob $1/3$
$S_k = \{2\}$ with prob $1/3$
$S_k = \{3\}$ with prob $1/3$

**Minibatch SAGA (with 2-nice sampling)**

$S_k = \{1, 2\}$ with prob $1/3$
$S_k = \{2, 3\}$ with prob $1/3$
$S_k = \{3, 1\}$ with prob $1/3$

**SAGA with nonuniform sampling**

$S_k = \{1\}$ with prob $p_1$
$S_k = \{2\}$ with prob $p_2$
$S_k = \{3\}$ with prob $p_3$
Arbitrary Sampling: Examples for $n = 3$

**GD**

$S_k = \{1, 2, 3\}$ with prob 1

**SAGA**

$S_k = \{1\}$ with prob $1/3$

$S_k = \{2\}$ with prob $1/3$

$S_k = \{3\}$ with prob $1/3$

**SAGA with nonuniform sampling**

$S_k = \{1\}$ with prob $p_1$

$S_k = \{2\}$ with prob $p_2$

$S_k = \{3\}$ with prob $p_3$

**Minibatch SAGA (with 2-nice sampling)**

$S_k = \{1, 2\}$ with prob $1/3$

$S_k = \{2, 3\}$ with prob $1/3$

$S_k = \{3, 1\}$ with prob $1/3$

**Interpolation between GD and SAGA**

$S_k = \{1, 2, 3\}$ with prob $1/2$

$S_k = \{1\}$ with prob $1/6$

$S_k = \{2\}$ with prob $1/6$

$S_k = \{3\}$ with prob $1/6$
A Brief History of Arbitrary Sampling
| #  | Paper                                                                 | Algorithm   | Comment                                                                                     |
|----|-----------------------------------------------------------------------|-------------|--------------------------------------------------------------------------------------------|
| 1  | R. & Takáč (OL 2016; arXiv 2013) On optimal probabilities in stochastic coordinate descent methods | NSync       | Arbitrary sampling (AS) first introduced Analysis of coordinate descent under strong convexity |
| 2  | Qu, R. & Zhang (NeurIPS 2015) Quartz: Randomized dual coordinate ascent with arbitrary sampling | QUARTZ      | First AS SGD method for min $P$ Primal-dual stochastic fixed point method; variance reduced   |
| 3  | Csiba & R. (arXiv 2015) Primal method for ERM with flexible mini-batching schemes and non-convex losses | Dual-free SDCA | First primal-only AS SGD method for min $P$ Variance-reduced                           |
| 4  | Qu & R. (OMS 2016) Coordinate descent with arbitrary sampling I: algorithms and complexity | ALPHA       | First accelerated coordinate descent method with AS Analysis for smooth convex functions     |
| 5  | Qu & R. (OMS 2016) Coordinate descent with arbitrary sampling II: expected separable overapproximation |             | First dedicated study of ESO inequalities needed for analysis of AS methods                   |
| 6  | Chambolle, Ehrhardt, R. & Schoenlieb (SIOPT 2018) Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications | SPDHGM      | Chambolle-Pock method with AS                                                               |
| 7  | Hanzely, Mishchenko & R. (NeurIPS 2018) SEGA: Variance reduction via gradient sketching | SEGA        | Variance-reduce coordinate descent with AS                                                  |
| 8  | Hanzely & R. (AISTATS 2019) Accelerated coordinate descent with arbitrary sampling and best rates for minibatches | ACD         | First accelerated coordinate descent method with AS Analysis for smooth strongly convex functions Importance sampling for minibatches |
| 9  | Horváth & R. (ICML 2019) Nonconvex variance reduced optimization with arbitrary sampling | SARAH, SVRG, SAGA | First non-convex analysis of an AS method First optimal mini-batch sampling                 |
| 10 | Gower, Loizou, Qian, Sailanbayev, Shulgin & R. (ICML 2019) SGD: general analysis and improved rates | SGD-AS      | First AS variant of SGD (without variance reduction) Optimal minibatch size                  |
| 11 | Qian, Qu & R. (ICML 2019) SAGA with arbitrary sampling                | SAGA-AS     | First AS variant of SAGA                                                                     |
The Algorithm
New Method: SAGA-AS (high level)

The Problem

\[
\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)
\]
New Method: SAGA-AS (high level)

Sample fresh $S_k \subseteq \{1, 2, \ldots, n\}$

The Problem

\[
\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)
\]
New Method: SAGA-AS (high level)

1. Sample fresh $S_k \subseteq \{1, 2, \ldots, n\}$

2. $J^{k+1}_{::i} = \begin{cases} \nabla f_i(x^k) & i \in S_k \\ J^k_{::i} & i \notin S_k \end{cases}$

The Problem

\[
\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)
\]

**Arbitrary Sampling**

**Jacobian Sketch**, i.e., a random matrix approximating the Jacobian:

\[
J^{k+1} \approx G(x^k) \overset{\text{def}}{=} [\nabla f_1(x^k), \ldots, \nabla f_n(x^k)] \in \mathbb{R}^{d \times n}
\]
New Method: SAGA-AS (high level)

1. Sample fresh $S_k \subseteq \{1, 2, \ldots, n\}$

2. $J_{k+1}^{i} = \begin{cases} \nabla f_i(x^k) & i \in S_k \\ J_{k}^{i} & i \notin S_k \end{cases}$

3. Use $J_{k+1}^{i}$, $J_{k}^{i}$ to build an unbiased estimator $g^k$ of $\nabla f(x^k)$

The Problem

$$\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)$$

Jacobian Sketch, i.e., a random matrix approximating the Jacobian:

$$J_{k+1}^{i} \approx G(x^k) \overset{\text{def}}{=} [\nabla f_1(x^k), \ldots, \nabla f_n(x^k)] \in \mathbb{R}^{d \times n}$$
**New Method:**

**SAGA-AS** (high level)

1. Sample fresh $S_k \subseteq \{1, 2, \ldots, n\}$

2. $J_{k+1}^i = \begin{cases} \nabla f_i(x^k) & i \in S_k \\ J_k^i & i \notin S_k \end{cases}$

3. Use $J_{k+1}$, $J_k$ to build an unbiased estimator $g^k$ of $\nabla f(x^k)$

4. $x^{k+1} = \text{prox}_{\alpha \psi} \left( x^k - \alpha g^k \right)$

---

**The Problem**

$$\min_{x \in \mathbb{R}^d} P(x) \overset{\text{def}}{=} \left( \sum_{i=1}^{n} \lambda_i f_i(x) \right) + \psi(x)$$

**Arbitrary Sampling**

**Jacobian Sketch**, i.e., a random matrix approximating the Jacobian:

$$J_{k+1} \approx G(x^k) \overset{\text{def}}{=} [\nabla f_1(x^k), \ldots, \nabla f_n(x^k)] \in \mathbb{R}^{d \times n}$$

**Proximal SGD step with fixed step size**

$$\text{prox}_{\psi}(x) \overset{\text{def}}{=} \arg \min_y \left\{ \frac{1}{2} \| y - x \|^2 + \psi(y) \right\}$$
Convergence Theory
## Convergence Theory

| Regime                  | Arbitrary sampling                                                                 | Thm |
|-------------------------|-------------------------------------------------------------------------------------|-----|
| **Smooth**              | \[
\begin{align*}
\psi & \equiv 0 \\
\text{f}_i \text{ is } L_i\text{-smooth, } f \text{ is } \mu\text{-strongly convex}
\end{align*}
\] | max \[
\max_{1 \leq i \leq n} \left\{ \frac{1}{p_i} + \frac{4(1 + B) L_i A_i \lambda_i}{\mu}, \frac{2B(1 + 1/B)L}{\mu} \right\} \log \left( \frac{1}{\epsilon} \right)
\] | 3.3 |
| **Nonsmooth**           | \[
\begin{align*}
P & \text{satisfies } \mu\text{-growth condition (19) and Assumption 4.3} \\
f_i(x) & = \phi_i(A_i^T x), \phi_i \text{ is } 1/\gamma\text{-smooth, } f \text{ is } L\text{-smooth}
\end{align*}
\] | \[
2 + \max_{1 \leq i \leq n} \left\{ \frac{6L}{\mu}, 3 \max_{1 \leq i \leq n} \left\{ \frac{1}{p_i} + \frac{4v_i \lambda_i}{p_i \mu \gamma} \right\} \right\} \log \left( \frac{1}{\epsilon} \right)
\] | 4.4 |
| **Nonsmooth**           | \[
\begin{align*}
\psi & \text{ is } \mu\text{-strongly convex} \\
f_i(x) & = \phi_i(A_i^T x), \phi_i \text{ is } 1/\gamma\text{-smooth}
\end{align*}
\] | max \[
\max_{1 \leq i \leq n} \left\{ 1 + \frac{1}{p_i} + \frac{3v_i \lambda_i}{p_i \mu \gamma} \right\} \log \left( \frac{1}{\epsilon} \right)
\] | 4.5 |

*Table 1.* Iteration complexity results for SAGA-AS. We have \( p_i := \mathbb{P}(i \in S) \), where \( S \) is a sampling of subsets of \([n]\) utilized by SAGA-AS. The key complexity parameters \( A_i, B, \) and \( v_i \) are defined in the sections containing the theorems.

**Expected Separable Over-approximation (ESO):**

\[
\mathbb{E}_S \left[ \left\| \sum_{i \in S} A_i h_i \right\|^2 \right] \leq \sum_{i=1}^{n} p_i v_i \| h_i \|^2
\]

\( p_i \overset{\text{def}}{=} \text{Prob}(i \in S_k) \)
Contributions
|                          | SAGA (Defazio et al 2014) | QUARTZ (Qu et al 2015) | JacSketch (Gower et al 2018) | SAGA-AS (THIS WORK) |
|--------------------------|---------------------------|------------------------|-----------------------------|-------------------|
| **PRIMAL / DUAL**        | Primal                    | Primal-dual            | Primal                      | Primal            |
| **SAMPLING**             | Uniform sampling of single data points | Arbitrary sampling (first AS method for min P) | A general sketching mechanism, but does not cover arbitrary sampling | Arbitrary sampling |
| **IMPORTANCE SAMPLING?** | NO                        | YES                    | YES (first SAGA-IS, but not for minibatches) | YES (also for minibatches) |
| **REGULARIZER**          | Support for any convex regularizer | Support for strongly convex regularizer | No support for a regularizer | Support for any convex regularizer |
| **RATE**                 | Linear                    | Linear                 | Linear                      | Linear (same or better) |
| **ASSUMPTIONS**          | Each $f_i$ strongly convex | strongly convex regularizer | Each $f_i$ strongly convex | $P$ satisfying quadratic growth |
| **HANDLING BIAS**        | Scaling                   | Built in               | Bias-correcting random variable | Bias-correcting random vector |
Experiments
**SDCA vs SAGA**

![Graph showing SDCA vs SAGA](image1)

![Graph showing SDCA vs SAGA](image2)
Uniform vs Importance Sampling

- ijcnn1
- w8a
What’s Next?
where $L_y$ is a smooth and convex, $A_i$ is twice differentiable, and $v^T(x-x^*)$ is closed and convex.

### Sampling

A random set valued mapping $F$ with online being subsets of $[1, \ldots, n]$. A sampling is uniquely defined by sampling probabilities $p_i$. Let $\tau \subseteq \{1, \ldots, n\}$ denote the sampling set.

The Expectations are computed with respect to $F$.

**Assumptions:**
- $f(x, y) = f(x) + y^T A_i x$ is smooth and convex.
- $A_i$ is twice differentiable.
- $v^T(x-x^*)$ is closed and convex.

### Algorithm

**Step 1 (Sampling):**

- Let $\tau$ be a properly chosen random set.
- Let $\tau = \emptyset$.

**Step 2 (Updating):**

- For $i = 1, 2, \ldots,$
- Sample $k_i \sim P_i$.
- Compute $x_{k_i} = \arg\min_x f(x, y) + \frac{\beta}{2} \|x - k_i\|^2$.
- Update $x_k$ using a method of choice.

**Step 3 (Termination):**

- Stop when $\|x_k - x^{k-1}\| < \epsilon$.

**References:**

[1] Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 18:483–514, 2017.

[2] Zheng Qu, Peter Richtárik, and Tong Zhang. Linear speedup in stochastic gradient descent. In *Advances in Neural Information Processing Systems 29*, pages 1662–1670. 2016.
The End