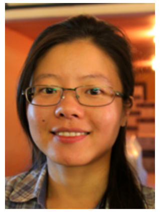
ICML | 2019

Thirty-sixth International Conference on Machine Learning

SAGA with Arbitrary Sampling


Xun Qian


Zheng Qu


Peter Richtárik

The Problem

The Problem: Regularized Empirical Risk Minimization

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

The Problem: Regularized Empirical Risk Minimization

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

The Problem: Regularized Empirical Risk Minimization

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

Regularizer

The Problem: Regularized Empirical Risk Minimization

training data

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

Regularizer

The Problem: Regularized Empirical Risk Minimization

training data

Weight associated with data point *i*

Loss associated with data point *i*

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

Regularizer

The Problem: Regularized Empirical Risk Minimization

training data

Weight associated with data point *i*

Loss associated with data point *i*

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

Parameters describing the model

Regularizer

Arbitrary Sampling

SGD with Arbitrary Sampling

SGD with Arbitrary Sampling

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

SGD with Arbitrary Sampling

1. In iteration k, we have x^k available
2. Sample a random set $S_k \subseteq \{1, 2, \ldots, n\}$

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

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

$$p_i \stackrel{\text{def}}{=} \mathrm{Prob}(i \in S_k)$$

$$p_i > 0 \text{ for all } i = 1, 2, \ldots, n$$

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: Examples for *n* = 3

GD

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

Arbitrary Sampling: Examples for *n* = 3

GD

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

SAGA

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

Arbitrary Sampling: Examples for $n = 3$

GD

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

SAGA

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

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

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

SAGA

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

SAGA with nonuniform sampling

$$S_k = \{1\} \text{ with prob } p_1$$
$$S_k = \{2\} \text{ with prob } p_2$$
$$S_k = \{3\} \text{ with prob } p_3$$

Minibatch SAGA (with 2-nice sampling)

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

Arbitrary Sampling: Examples for $n = 3$

GD

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

SAGA

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

SAGA with nonuniform sampling

$$S_k = \{1\} \text{ with prob } p_1$$
$$S_k = \{2\} \text{ with prob } p_2$$
$$S_k = \{3\} \text{ with prob } p_3$$

Minibatch SAGA (with 2-nice sampling)

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

Interpolation between GD and SAGA

$$S_k = \{1,2,3\} \text{ with prob } 1/2$$
$$S_k = \{1\} \text{ with prob } 1/6$$
$$S_k = \{2\} \text{ with prob } 1/6$$
$$S_k = \{3\} \text{ with prob } 1/6$$

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

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

New Method: SAGA-AS (high level)

The Problem

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

Arbitrary Sampling

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

New Method: SAGA-AS (high level)

The Problem

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

Arbitrary Sampling

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

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

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

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

New Method: SAGA-AS (high level)

The Problem

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

Arbitrary Sampling

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

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

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

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

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

New Method: SAGA-AS (high level)

The Problem

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

Arbitrary Sampling

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

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

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

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

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

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

Proximal SGD step with fixed step size

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

Convergence Theory

Convergence Theory

$$\mathbb{E}_S\left[\left\|\mathbf{M}\,\mathrm{Diag}\left(\theta_S\right)\mathbf{I}_S\lambda\right\|^2\right] \le \sum_{i=1}^n \mathcal{A}_i\lambda_i^2\left\|\mathbf{M}_{:i}\right\|^2 + \mathcal{B}\|\mathbf{M}\lambda\|^2$$

Lyapunov function:

$$\Psi^k \overset{\text{def}}{=} \left\|x^k - x^*\right\|^2 + 2\alpha\sum_{i=1}^n \sigma_i\mathcal{A}_i\lambda_i^2\left\|\mathbf{J}_{:i}^k - \nabla f_i\left(x^*\right)\right\|^2$$

Regime	Arbitrary sampling	Thm
Smooth $\psi \equiv 0$ f_i is L_i-smooth, f is μ-strongly convex	$\max\left\{\max_{1\le i\le n}\left\{\frac{1}{p_i}+\frac{4(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i}{\mu}\right\}, \frac{2\mathcal{B}(1+1/\mathcal{B})L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right)$	3.3
Nonsmooth P satisfies μ-growth condition (19) and Assumption 4.3 $f_i(x) = \phi_i(\mathbf{A}_i^\top x)$, ϕ_i is $1/\gamma$-smooth, f is L-smooth	$\left(2+\max\left\{\frac{6L}{\mu}, 3\max_{1\le i\le n}\left\{\frac{1}{p_i}+\frac{4v_i\lambda_i}{p_i\mu\gamma}\right\}\right\}\right)\log\left(\frac{1}{\epsilon}\right)$	4.4
Nonsmooth ψ is μ-strongly convex $f_i(x) = \phi_i(\mathbf{A}_i^\top x)$, ϕ_i is $1/\gamma$-smooth	$\max_{1\le i\le n}\left\{1+\frac{1}{p_i}+\frac{3v_i\lambda_i}{p_i\mu\gamma}\right\}\log\left(\frac{1}{\epsilon}\right)$	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 $\mathcal{A}_i$, $\mathcal{B}$, and v_i are defined in the sections containing the theorems.

Expected Separable Over-approximation (ESO):

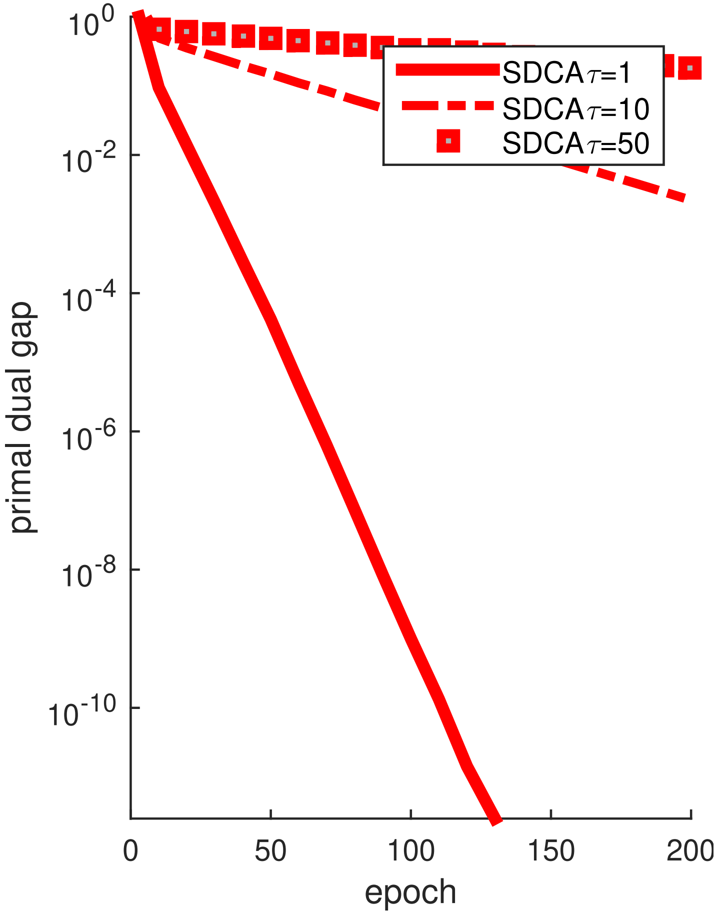
$$\mathbb{E}_S\left[\left\|\sum_{i\in S}\mathbf{A}_i h_i\right\|^2\right] \le \sum_{i=1}^n p_i v_i\left\|h_i\right\|^2 \qquad p_i \overset{\text{def}}{=} \mathrm{Prob}(i \in S_k)$$

Contributions

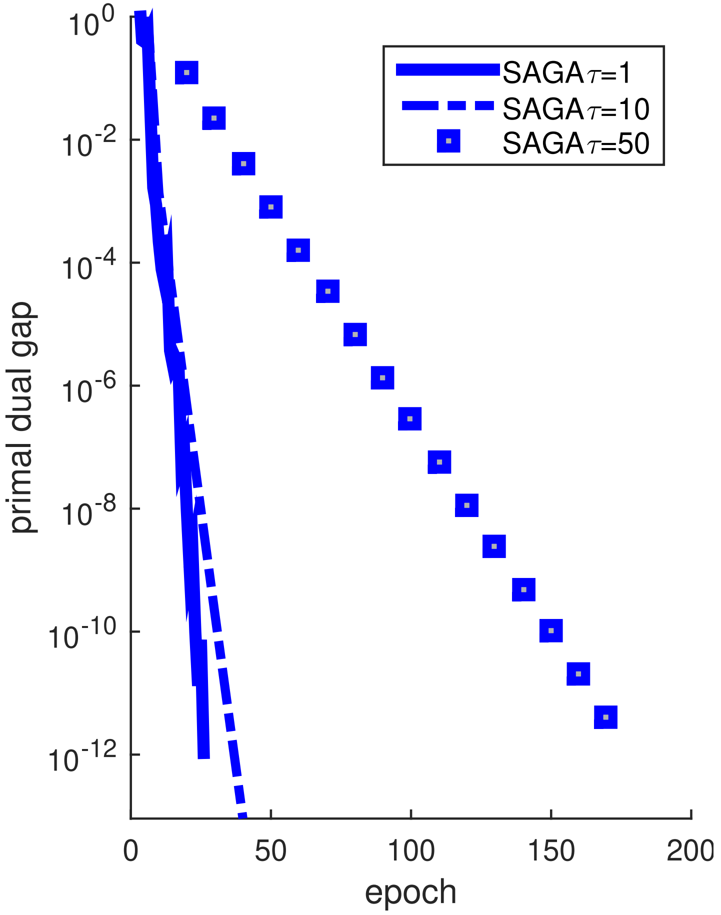
	SAGA (Defazio et al 2014)	QUARTZ (Qu et al 2015)	JacSketch (Gower et al 2018)	SAGA-AS (THIS WORK)
PRIMAL / DUAL	Primal	Primal-dual	Primal	Primal
SAMPLING	Uniform sampling of of single data points	Arbitrary sampling (first AS method for min P)	A general sketching mechanism, but does not cover arbitrary 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

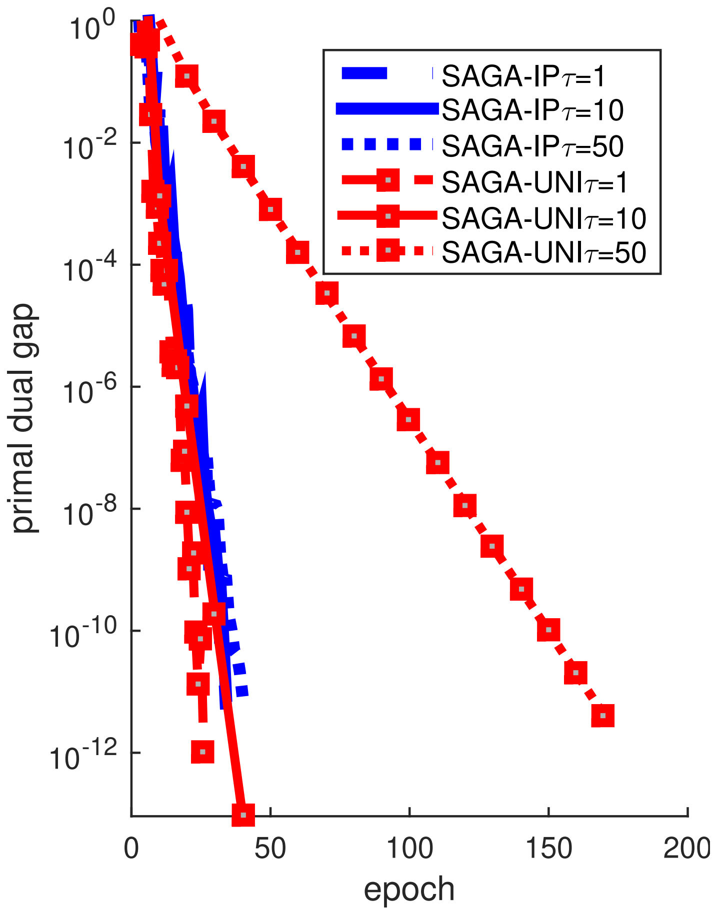


ijcnn1

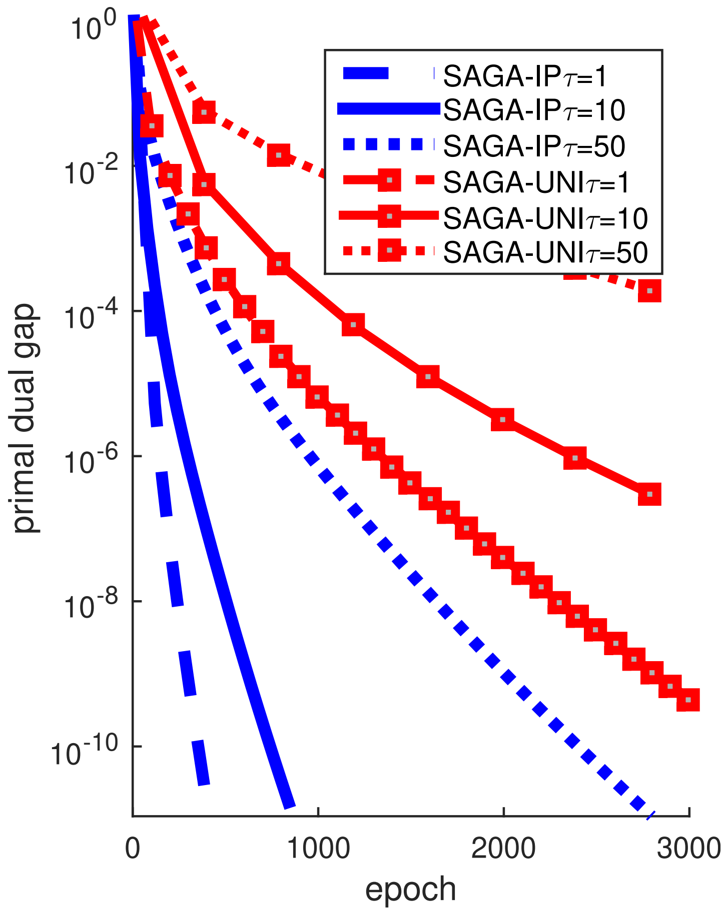


ijcnn1

Uniform vs Importance Sampling



ijcnn1



w8a

What’s Next?

SAGA with Arbitrary Sampling

Xun Qian[1] Zheng Qu [2] Peter Richtárik[1, 3, 4]

[1]KAUST [2]University of Hong Kong [3]University of Edinburgh [4]Moscow Institute of Physics and Technology

The Problem

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

where $f \stackrel{\text{def}}{=} \sum_{i=1}^n \lambda_i f_i(x)$, f_i are smooth and convex, $\lambda_i > 0$ are weights, and $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is closed and convex.

Sampling

Sampling: A random set valued mapping S with values being subsets of $\{1,\ldots,n\}$. A sampling is uniquely defined by assigning probabilities to all 2^n subsets of $\{1,\ldots,n\}$. Let $\tau \stackrel{\text{def}}{=} \mathbb{E}|S|$ be the expected size of S, and define

$$p_i \stackrel{\text{def}}{=} \text{Prob}(i \in S), \quad i \in \{1,\ldots,n\}.$$

A sampling is called proper if $p_i > 0$ for all i. For $C \subseteq \{1,\ldots,n\}$, let

$$p_C \stackrel{\text{def}}{=} \text{Prob}(S = C).$$

Bias-correcting random vector: vector $\theta_S = (\theta_S^1,\ldots,\theta_S^n) \in \mathbb{R}^n$ with the property

$$\mathbb{E}\left[\text{Diag}(\theta_S)\mathbf{I}_S e\right] = e, \quad \text{i.e.,} \quad \mathbb{E}\left[\theta_S^i 1_{i\in S}\right] = 1, \ \forall i, \quad (2)$$

where

- e: $n \times 1$ vector of all ones
- $\mathbf{I}$: $n \times n$ identity matrix
- $\mathbf{I}_S$: $n \times n$ matrix with ones in places (i,i) for $i \in S$
- $1_{i\in S}$: indicator random variable of the event $i \in S$, i.e.,: $1_{i\in S} = 1$ if $i \in S$ and $1_{i\in S} = 0$ if $i \notin S$

Algorithm

Prox operator: $\text{prox}_\alpha^\psi(x) \stackrel{\text{def}}{=} \arg\min\left\{\frac{1}{2\alpha}\|x-y\|^2 + \cdots\right.$

Gradient matrix: $\mathbf{G}(x) \stackrel{\text{def}}{=} [\nabla f_1(x), \cdots, \nabla$

Algorithm 1: SAGA with Arbitr

Initialize: $x^0 \in \mathbb{R}^d$, $\mathbf{J}^0 \in$

Parameters: arbit

θ_S, stepsize $\alpha >$

for $k = 1, 2, \ldots$

Sample fresh $S_k \subseteq$

$\mathbf{J}^{k+1} = \mathbf{J}^k + (\mathbf{G}(x^k$

$g^k = \mathbf{J}^k\lambda + (\mathbf{G}(x^k) -$

$x^{k+1} = \text{prox}_\alpha^\psi(x^k - \alpha g$

end

Smoo

Assumptions:

- f_i is convex and L_i-smooth,
- f is μ-strongly convex and L-sm
- There exist constants $\mathcal{A}_i \geq 0$ and $0 \leq \mathcal{B} \leq 1$ such that for any matrix $\mathbf{M} \in \mathbb{R}^{d\times n}$

$$\mathbb{E}\left[\|\mathbf{M}\text{Diag}(\theta_S)\mathbf{I}_S\lambda\|^2\right] \leq \sum_{i=1}^n \mathcal{A}_i\lambda_i^2\|\mathbf{M}_{:i}\|^2 + \mathcal{B}\|\mathbf{M}\lambda\|^2$$

Lyapunov function:

$$\Psi^k \stackrel{\text{def}}{=} \|x^k - x^*\|^2 + 2\alpha\sum_{i=1}^n \sigma_i\mathcal{A}_i\lambda_i^2\|\mathbf{J}_{:i}^k - \nabla f_i(x^*)\|^2,$$

where $\sigma_i = \frac{1}{4(1+\mathcal{B})L_i\mathcal{A}_i p_i\lambda_i}$ and x^* is a solution of (1).

Convergence Result $(\mathbb{E}[\Psi^k] \leq \epsilon \cdot \mathbb{E}[\Psi^0])$

μ is known: $\alpha = \min_i\left\{\frac{p_i}{\mu+4(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i p_i}, \frac{\mathcal{B}^{-1}}{2(1+1/\mathcal{B})L}\right\}$

$$k \geq \max_i\left\{\frac{1}{p_i} + \frac{4(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i}{\mu}, \frac{2\mathcal{B}(1+\frac{1}{\mathcal{B}})L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right).$$

μ is unknown: $\alpha = \min_i\left\{\frac{p_i}{8(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i p_i}, \frac{\mathcal{B}^{-1}}{2(1+1/\mathcal{B})L}\right\}$

$$k \geq \max_i\left\{\frac{2}{p_i}, \frac{8(1+\mathcal{B})L_i\mathcal{A}_i\lambda_i}{\mu}, \frac{2\mathcal{B}(1+\frac{1}{\mathcal{B}})L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right).$$

Interface For Sampling

- Proper sampling: $\mathcal{A}_i = \beta_i \stackrel{\text{def}}{=} \sum_{C\subseteq[n]:i\in C} p_C|C|(\theta_C^i)^2$, $\mathcal{B} = 0$.
- τ-nice sampling ($\theta_S^i = \frac{1}{p_i}$): $\mathcal{A}_i = \frac{n}{\tau}\cdot\frac{n-\tau}{n-1}$, $\mathcal{B} = \frac{n(\tau-1)}{\tau(n-1)}$.
- Independent sampling ($\theta_S^i = \frac{1}{p_i}$): $\mathcal{A}_i = \frac{1}{p_i} - 1$, $\mathcal{B} = 1$.

Optimal Bias-Correcting Random Vector

Let $\Theta(S)$ be the collection of all bias-correcting random ated with sampling S, i.e., $\mathbb{E}[\theta_S\mathbf{I}_S e] = e$. Let

Lo

Let S be a proper

ance Sampling

e the expected minibatch size, and $\bar{L} \stackrel{\text{def}}{=} \sum_{i\in[n]} L_i\lambda_i$.
e independent sampling with $\theta_S^i = 1/p_i$. Let

$$q_i = \frac{(\mu + 8L_i\lambda_i)\tau}{\sum_{i\in[n]}(\mu + 8L_i\lambda_i)}.$$

By choosing $\min\{q_i, 1\} \leq p_i \leq 1$ such that $\sum_{i\in[n]} p_i = \tau$, the iteration complexity becomes:

$$\max\left\{\frac{n}{\tau} + \frac{8\bar{L}}{\mu\tau}, \frac{4L}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right). \quad (3)$$

Linear speedup: When $\tau \leq \frac{n\mu+8\bar{L}}{4L}$, (3) becomes

$$\left(\frac{n}{\tau} + \frac{8\bar{L}}{\mu\tau}\right)\log\left(\frac{1}{\epsilon}\right),$$

which yields linear speedup with respect to τ. When $\tau \geq \frac{n\mu+8\bar{L}}{4L}$, (3) becomes

$$\frac{4L}{\mu}\log\left(\frac{1}{\epsilon}\right).$$

Nonsmooth Case (strongly convex)

Assumptions:

- $f_i(x) = \phi_i(\mathbf{A}_i^\top x)$
- ϕ is $1/\gamma$-smooth and convex
- ψ_i is μ-strongly convex
- Choose $\theta_S^i = 1/p_i$
- Let v_i satisfy the ESO inequality:

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

Lyapunov function:

$$\Psi^k \stackrel{\text{def}}{=} \|x^k - x^*\|^2 + \alpha\sum^n$$

C

ongly convex)

th and convex

$/p_i$

- ESO inequality
- Nullspace consistency: For any $x^*, y^* \in \mathcal{X}^*$ we have

$$\mathbf{A}_i^\top x^* = \mathbf{A}_i^\top y^*, \quad \forall i \in [n],$$

where $\mathcal{X}^* \stackrel{\text{def}}{=} \arg\min\{P(x) : x \in \mathbb{R}^d\}$.

- Quadratic functional growth condition: there is a constant $\mu > 0$ such that

$$P(x^k) - P^* \geq \frac{\mu}{2}\|x^k - [x^k]^*\|^2, \ w.p.1, \ \forall k \geq 1,$$

where $[x]^* = \arg\min\{\|x-y\| : y \in \mathcal{X}^*\}$, for the sequence $\{x^k\}$ produced by the Algorithm.

Lyapunov function:

$$\Psi^k \stackrel{\text{def}}{=} \|x^k - [x^k]^*\|^2 + \alpha\sum_{i=1}^n \sigma_i\frac{v_i}{p_i}\lambda_i^2\|\alpha_i^k - \nabla\phi_i(\mathbf{A}_i^\top x^*)\|^2,$$

where $\sigma_i = \gamma/2v_i\lambda_i$.

Convergence Result $(\mathbb{E}[\Psi^k] \leq \epsilon \cdot \mathbb{E}[\Psi^0])$

μ is known: $\alpha = \min\left\{\frac{2}{3}\min_{1\leq i\leq n}\frac{p_i}{\mu+4v_i\lambda_i/\gamma}, \frac{1}{3L}\right\}$

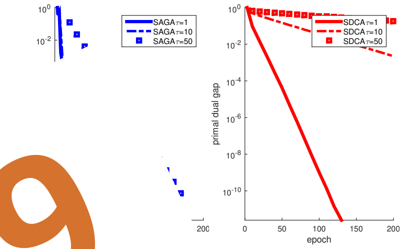
$$k \geq \left(2 + \max\left\{\frac{6L}{\mu}, 3\max_i\left(\frac{1}{p_i} + \frac{4v_i\lambda_i}{p_i\mu\gamma}\right)\right\}\right)\log\left(\frac{1}{\epsilon}\right).$$

μ is unknown: $\alpha = \min\left\{\min_{1\leq i\leq n}\frac{p_i}{12v_i\lambda_i/\gamma}, \frac{1}{3L}\right\}$

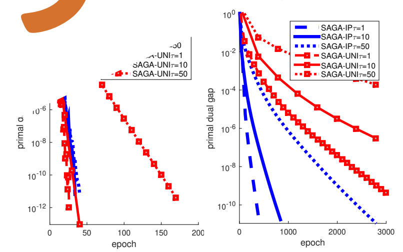
$$k \geq \left(2 + \max\left\{\frac{6L}{\mu}, \max_i\left\{\frac{24v_i\lambda_i}{\mu p_i\gamma}, \frac{2}{p_i}\right\}\right\}\right)\log\left(\frac{1}{\epsilon}\right).$$

Numerical Results

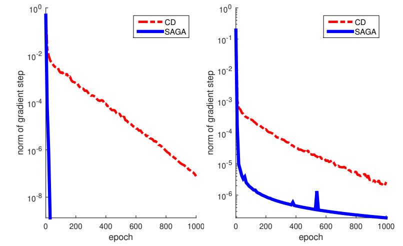
1. mini-batch SAGA versus mini-batch SDCA [1, 2]



rm sampling



3. SAGA versus CD



References

[1] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

[2] Zheng Qu, Peter Richtárik, and Tong Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *Advances in Neural Information Processing Systems 28*, pages 865–873. Curran Associates, Inc., 2015.

[3] R. M. Gower, P. Richtárik, and F. Bach. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *arXiv Preprint arXiv: 1805.02632*, 2018.

[4] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27*, 2014.

The End