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

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

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 $\{1,\dots,n\}$. A sampling is uniquely defined by assigning probabilities to all 2^n subsets of $\{1,\dots,n\}$. Let $\tau \stackrel{\text{def}}{=} \mathbb{E}|S|$ be the expected size of S, and define

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

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

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

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

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

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

Gradient matrix: $\mathbf{G}(x) \stackrel{\text{def}}{=} [\nabla f_1(x),\cdots,\nabla f_n(x)] \in \mathbb{R}^{d\times n}$

Algorithm 1: `SAGA` with Arbitrary Sampling (`SAGA-AS`)

Initialize: $x^0\in\mathbb{R}^d$, $\mathbf{J}^0\in\mathbb{R}^{d\times n}$
Parameters: arbitrary sampling S, bias-correcting random vector θ_S, stepsize $\alpha > 0$
for $k = 1,2,\dots$ **do**
  Sample fresh $S_k \subseteq \{1,\dots,n\}$
  $\mathbf{J}^{k+1} = \mathbf{J}^k + (\mathbf{G}(x^k) - \mathbf{J}^k)\mathbf{I}_{S_k}$
  $g^k = \mathbf{J}^k\lambda + (\mathbf{G}(x^k)-\mathbf{J}^k)\text{Diag}(\theta_{S_k})\mathbf{I}_{S_k}\lambda$
  $x^{k+1} = \text{prox}_\alpha^\psi(x^k - \alpha g^k)$
end

Smooth Case ($\psi \equiv 0$)

Assumptions:

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

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

Lyapunov function:

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

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

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

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

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

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

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 vectors associated with sampling S, i.e., $\mathbb{E}[\theta_S\mathbf{I}_S e] = e$. Let $\mathbb{E}^i[\cdot] \stackrel{\text{def}}{=} \mathbb{E}[\cdot \mid i\in S]$.

Lemma

Let S be a proper sampling. Then

- $$\min_{\theta\in\Theta(S)} \beta_i = \frac{1}{\sum_{C:i\in C} p_C/|C|} = \frac{1}{p_i\mathbb{E}^i[1/|S|]}$$

  for all i, and the minimum is obtained at $\theta\in\Theta(S)$ given by

  $$\theta_C^i = \frac{1}{|C|\sum_{C:i\in C}p_C/|C|} = \frac{1}{p_i|C|\mathbb{E}^i[1/|S|]}$$

  for all $C : i\in C$;
- Moreover,

  $$\frac{1}{\mathbb{E}^i[1/|S|]} \le \mathbb{E}^i[|S|], \quad \forall i\in\{1,\dots,n\}.$$

Importance Sampling

Let $\tau \stackrel{\text{def}}{=} \mathbb{E}[|S|]$ be the expected minibatch size, and $\bar{L} \stackrel{\text{def}}{=} \sum_{i\in[n]} L_i\lambda_i$. Consider the 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 $\min\{q_i, 1\} \le p_i \le 1$ such that $\sum_{i\in[n]} p_i = \tau$, the iteration complexity becomes:

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

Linear speedup: When $\tau \le \frac{n\mu + 8\bar{L}}{4L}$, (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 $\tau \ge \frac{n\mu+8\bar{L}}{4L}$, (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:

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

Lyapunov function:

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

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

μ **is known:** $\sigma_i = 2\gamma/3v_i\lambda_i$, $\alpha = \min_{1\le i\le n}\frac{p_i}{\mu+3v_i\lambda_i/\gamma}$

$$k \ge \max_i\left\{1 + \frac{1}{p_i} + \frac{3v_i\lambda_i}{p_i\mu\gamma}\right\}\log\left(\frac{1}{\epsilon}\right).$$

μ **is unknown:** $\sigma_i = \gamma/(1+\alpha\mu)v_i\lambda_i$, $\alpha = \min_{1\le i\le n}\frac{p_i\gamma}{4v_i\lambda_i}$

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

Nonsmooth Case (non-strongly convex)

Assumptions:

- $f_i(x) = \phi_i(\mathbf{A}_i^\top x)$
- ϕ is $1/\gamma$-smooth and convex
- $\theta_S^i = 1/p_i$
- 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

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

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

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

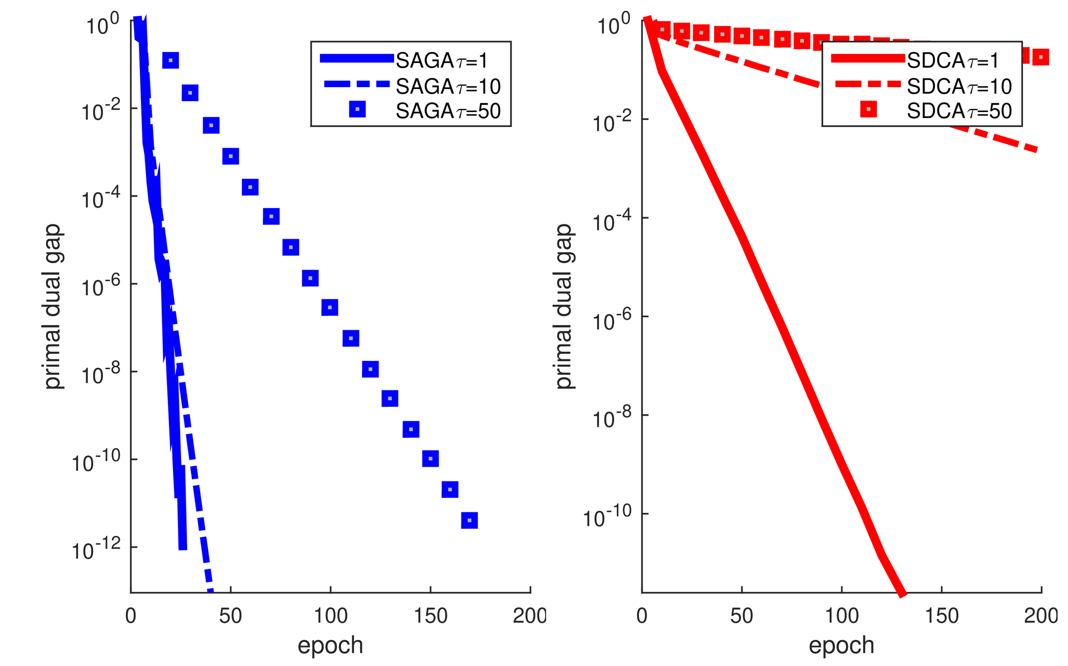
$$k \ge \left(2 + \max\left\{\frac{6L}{\mu}, 3\max_i\left(\frac{1}{p_i} + \frac{4v_i\lambda_i}{p_i\mu\gamma}\right)\right\}\right)\log\left(\frac{1}{\epsilon}\right).$$

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

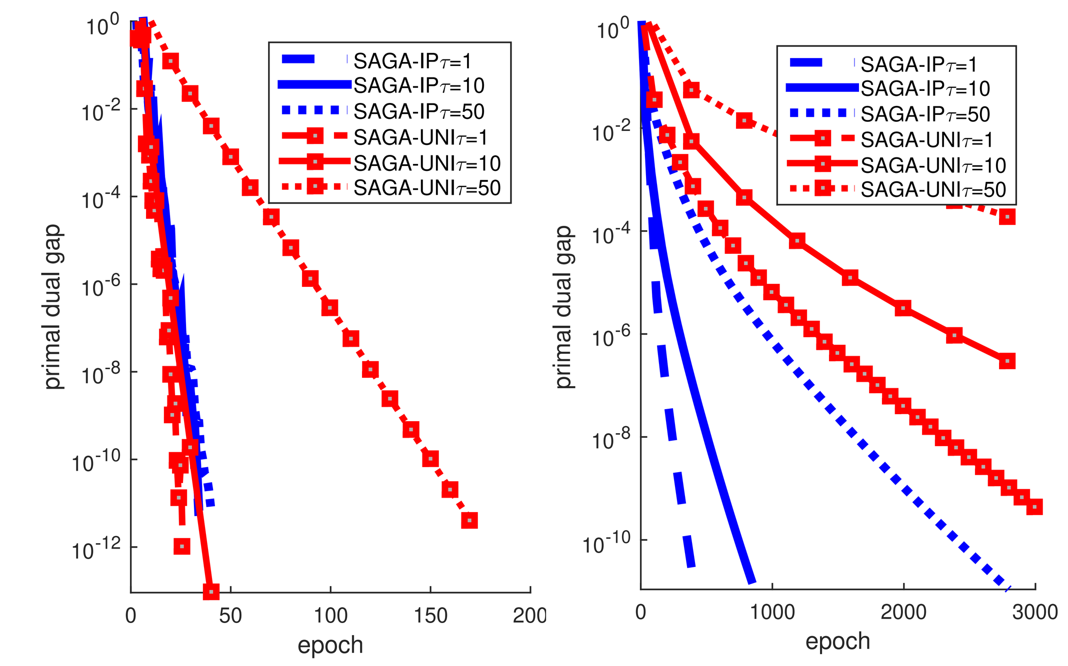
$$k \ge \left(2 + \max\left\{\frac{6L}{\mu}, \max_i\left\{\frac{24v_i\lambda_i}{\mu p_i\gamma}, \frac{2}{p_i}\right\}\right\}\right)\log\left(\frac{1}{\epsilon}\right).$$

Numerical Results

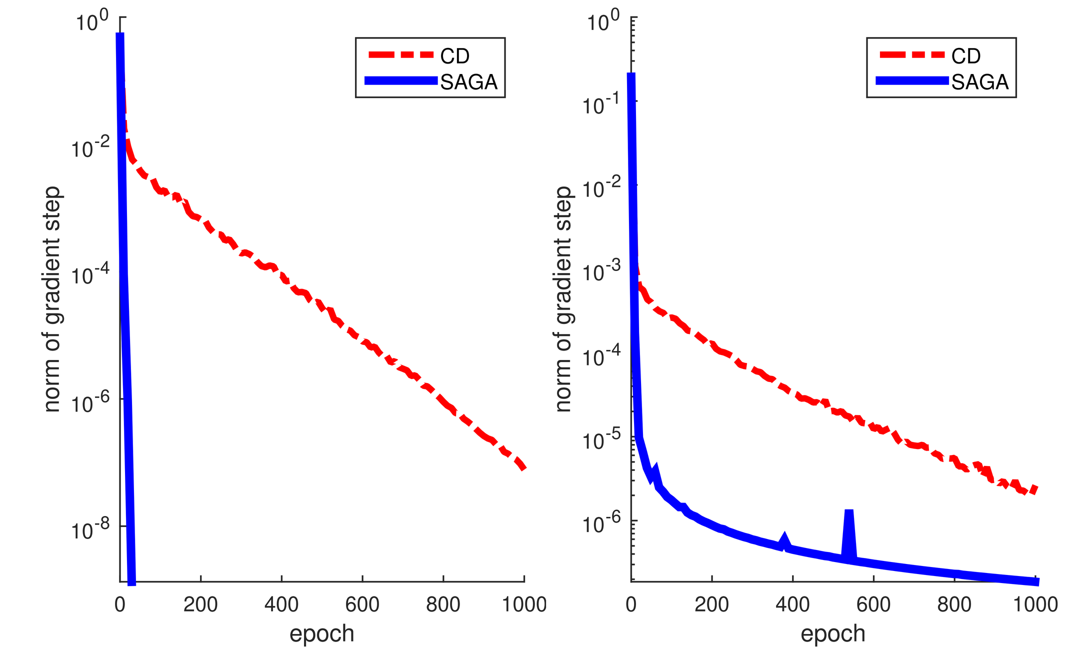
1. mini-batch SAGA versus mini-batch SDCA [1, 2]



2. Importance sampling versus uniform 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.