An Accelerated Variance Reducing Stochastic Method with Douglas-Rachford Splitting

Jingchang Liu

November 12, 2018

University of Science and Technology of China

Table of Contents

Background

Moreau Envelop and Douglas-Rachford (DR) Splitting

Our methods

Theories

Experiments

Conclusions

Q & A

Background

Problem

Formulation

- Regularized ERM: $\min_{x \in \mathcal{R}^d} f(x) + h(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h(x)$.
- $f_i : \mathbb{R}^d \to \mathbb{R}$: empirical loss of i-th sample, convex.
- h: regularization term, convex but possibly non-smooth.
- Examples: LASSO, sparse SVM, ℓ_1, ℓ_2-Logistic Regression.

Definition

- Proximal operator: $\text{prox}_f^\gamma(x) = \text{argmin}_{y \in \mathbb{R}^d} \left(f(y) + \frac{1}{2\gamma} \|y - x\|^2 \right)$.
- Gradient mapping: $f(x) = \frac{1}{\gamma}(x - \text{prox}_f^\gamma(x))$.
- Subdifferential: $\partial f(x) = \left\{ g \mid g^T(y - x) \leq f(y) - f(x), \forall y \in \text{dom } f \right\}$.
- Strongly convex: $f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\mu}{2} \|y - x\|^2$.
- L-smooth: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$.

Related Works

Exsiting Algorithm

$\text{prox}_h^\gamma(x - \gamma \cdot \square)$, where $\square$ can be obtained from:

- GD: $\square = \nabla f(x)$, more calculations needed in each iteration.
- SGD: $\square = \nabla f_i(x)$, small stepsize deduces slow convergence.
- Variance reduction (VR): $\square = \nabla f_i(x) - \nabla f_i(\bar{x}) + \nabla f(x)$, such as SVRG, SAGA, SDCA.

Accelerated Technique

- Ill condition: L/μ, the condition number, is large.
- Methods: Acc-SDCA, Catalyst, Mig, Point-SAGA.
- Drawbacks: More parameters need to be tuned.

Rate

Convergence Rate

- VR stochastic methods: $\mathcal{O}\left((n + L/\mu)\log(1/\epsilon)\right)$.
- Acc-SDCA, Mig, Point-SAGA: $\mathcal{O}((n + \sqrt{nL/\mu})\log(1/\epsilon))$.
- When $L/\mu \gg n$, accelerated technique makes the convergence much faster.

Aim

Design a simpler accelerate VR stochastic method which can achieve the fastest convergence rate.

Moreau Envelop and Douglas-Rachford (DR) Splitting

Moreau Envelop

Formulaton

$$f^{\gamma}(x) = \inf_{y} \left\{ f(y) + \frac{1}{2\gamma} \|x - y\|^2 \right\}.$$

Properties

- x^* minimizes $f(x)$ iff x^* minimizes $f^{\gamma}(x)$
- f^{γ} is continuously differentiable even when f is non-differentiable,

  $$\nabla f^{\gamma}(x) = (x - \mathrm{prox}_f^{\gamma}(x))/\gamma.$$

  Moreover, f^{γ} is $1/\gamma$-smooth.
- If f: μ-strongly convex, then f^{γ}: $\mu/(\mu\gamma + 1)$-strongly convex.
- The condition number of f^{γ} is $(\mu\gamma + 1)/\mu\gamma$, which may be better.

Proximal Point Algorithm (PPA)

$$x^{k+1} = \mathrm{prox}_f^{\gamma}(x^k) = x^k - \gamma \nabla f^{\gamma}(x^k).$$

Point-SAGA

Formulation

Used when h is absent: $\min_{x \in \mathcal{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$

Iteration

$$\begin{aligned} z_j^k &= x^k + \gamma(g_j^k - \sum_{i=1}^{n} g_i^k / n), \\ x^{k+1} &= \operatorname{prox}_{f_j}^{\gamma}(z_j^k) \\ g_j^{k+1} &= (z_j^k - x^{k+1})/\gamma, \end{aligned}$$

Equivalence

$$x^{k+1} = x^k - \gamma\big(g_j^{k+1} - g_j^k + \sum_{i=1}^{n} g_i^k / n\big),$$

where g_j^{k+1} is the gradient mapping of f at z_j^k.

Point-SAGA: Convergence rate

Strongly convex and smooth

$$\mathcal{O}\left((n + \sqrt{n\frac{L}{\mu}}) \log(\frac{1}{\epsilon})\right).$$

Strongly convex and non-smooth

$$\mathcal{O}\left(\frac{1}{\epsilon}\right).$$

Douglas-Rachford (DR) Splitting

Formulation

$$\min_{x \in \mathbb{R}^d} f(x) + h(x),$$

Iteration

$$\begin{aligned} y^{k+1} &= -x^k + y^k + \mathrm{prox}_f^\gamma(2x^k - y^k), \\ x^{k+1} &= \mathrm{prox}_h^\gamma(y^{k+1}). \end{aligned}$$

Convergence

- $F(y) = y + \mathrm{prox}_h^\gamma(2\mathrm{prox}_f^\gamma(y) - y) - \mathrm{prox}_f^\gamma(y)$.
- y is a fixed point of F if and only if $x = \mathrm{prox}_f^\gamma(y)$ satisfies $0 \in \partial f(x) + \partial g(x)$:

$$y = F(y) \quad \rightleftarrows \quad 0 \in \partial f(\mathrm{prox}_y^\gamma(y)) + \partial g(\mathrm{prox}_y^\gamma(y)).$$

Our methods

Algorithm

Algorithm 1 Prox2-SAGA

Input: $x^0 \in \mathbb{R}^d$, g_i^0 $(i = 1, 2, \ldots, n)$, step size $\gamma > 0$.
for $k = 0, 1, \ldots$ **do**
    Uniformly randomly pick j from 1 to n.
    Calculate g_j^{k+1}:

$$
\begin{aligned}
z_j^k &= x^k + \gamma\Big(g_j^k - \frac{1}{n}\sum_{i=1}^{n} g_i^k\Big), && (8)\\
g_j^{k+1} &= \frac{1}{\gamma}\big((z_j^k + x^k - y^k) - \mathrm{prox}_{f_j}^{\gamma}(z_j^k + x^k - y^k)\big). && (9)
\end{aligned}
$$

    Update x:

$$
\begin{aligned}
y^{k+1} &= z_j^k - \gamma g_j^{k+1}, && (10)\\
x^{k+1} &= \mathrm{prox}_h^{\gamma}(y^{k+1}). && (11)
\end{aligned}
$$

    Update g_i $(i = 1, 2, \ldots, n)$ in the table:

$$
g_i^{k+1} = \begin{cases} g_j^{k+1}, & \text{if } i = j,\\ g_i^k, & \text{otherwise.} \end{cases} \qquad (12)
$$

end for
Output: x^{k+1}.

Iterations

Main iterations

$$\begin{aligned} y^{k+1} &= x^k - \gamma\Big(g_j^{k+1} - g_j^k + \frac{1}{n}\sum_{i=1}^n g_i^k\Big), \\ x^{k+1} &= \mathrm{prox}_h^\gamma(y^k), \end{aligned}$$

where

$$g_j^{k+1} = \frac{1}{\gamma}\big((z_j^k + x^k - y^k) - \mathrm{prox}_{f_j}(z_j^k + x^k - y^k)\big),$$

the gradient mapping of f_j at $z_j^k - x^k - y^k$.

Number of parameters

Prox2-SAGA	Point-SAGA	Katyusha	Mig	Acc-SDCA	Catalyst
1	1	3	2	2	several

Connections to other algorithms

Point-SAGA

When h $= 0$, we have $x_k = y_k$ for Prox2-SAGA,

$$\begin{aligned} z_j^k &= x^k + \gamma\Big(g_j^k - \frac{1}{n}\sum_{i=1}^n g_i^k\Big), \\ x^{k+1} &= \mathrm{prox}_{f_j}^{\gamma}(z_j^k), \\ g_j^{k+1} &= \frac{1}{\gamma}(z_j^k - x^{k+1}). \end{aligned}$$

DR splitting

When $n = 1$, since $g_j^k = \sum_{i=1}^n g_i^k/n$ in Prox2-SAGA,

$$\begin{aligned} y^{k+1} &= -x^k + y^k + \mathrm{prox}_f^{\gamma}(2x^k - y^k), \\ x^{k+1} &= \mathrm{prox}_h^{\gamma}(y^{k+1}). \end{aligned}$$

Theories

Effectiveness

Proposition
Suppose that $(y^\infty, \{g_i^\infty\}_{i=1,\dots,n})$ *is the fixed point of the Prox2-SAGA iteration. Then* $x^\infty = \mathrm{prox}_h^\gamma(y^\infty)$ *is a minimizer of* $f + h$*.*

Proof.
$\because y^\infty = -x^\infty + y^\infty + \mathrm{prox}_{f_i}^\gamma(z_i^\infty + x^\infty - y^\infty)$, which implies

$$(z_i^\infty - y^\infty)/\gamma \in \partial f_i(x^\infty),\ i = 1, \dots, n. \tag{1}$$

Meanwhile, because $x^\infty = \mathrm{prox}_h^\gamma(y^\infty)$, we have

$$(y^\infty - x^\infty)/\gamma \in \partial h(x^\infty). \tag{2}$$

Observing that

$$\frac{1}{n}\sum_{i=1}^{n}(z_i^\infty - y^\infty) + (y^\infty - x^\infty) = \frac{1}{n}\sum_{i=1}^{n} z_i^\infty - x^\infty = 0,$$

from (1) and (2), we have $0 \in \partial f(x^\infty) + \partial h(x^\infty)$. $\square$

Convergence Rate

Non-strongly convex case

Suppose that f_i: convex and L-smooth, h: convex. Denote $\bar{g}_j^k = \frac{1}{k}\sum_{t=1}^k g_j^t$, then for Prox2-SAGA with step size $\gamma \leq 1/L$, at any time $k > 0$ it holds

$$\mathbb{E}\left\|\bar{g}_j^k - g_j^*\right\|^2 \leq \frac{1}{k}\Big(\sum_{i=1}^n \left\|g_i^0 - g_i^*\right\|^2 + \|\frac{1}{\gamma}(y^0 - y^*)\|^2\Big).$$

Strongly convex case

Suppose that f_i: μ-strongly convex and L-smooth, h: convex. Then for Prox2-SAGA with stepsize $\gamma = \min\left\{\frac{1}{\mu n}, \frac{\sqrt{9L^2+3\mu L}-3L}{2\mu L}\right\}$, for any time $k > 0$ it holds

$$\mathbb{E}\left\|x^k - x^*\right\|^2 \leq \big(1 - \frac{\mu\gamma}{2\mu\gamma+2}\big)^k \cdot \frac{\mu\gamma - 2}{2 - n\mu\gamma}\Big\{\sum_{i=1}^n \left\|\gamma(g_i^0 - g_i^*)\right\|^2 + \|y^0 - y^*\|^2\Big\}.$$

Remarks

- When the stepsize

$$\gamma = \min\left\{\frac{1}{\mu n}, \frac{\sqrt{9L^2 + 3\mu L} - 3L}{2\mu L}\right\},$$

then $\mathcal{O}(n + L/\mu)\log(1/\epsilon)$ steps are required to achieve $\mathbb{E}\left\|x^k - x^*\right\|^2 \leq \epsilon$.

- When f_i is ill-conditioned, then a large stepsize

$$\gamma = \min\left\{\frac{1}{\mu n}, \frac{6L + \sqrt{36L^2 - 6(n-2)\mu L}}{2(n-2)\mu L}\right\}$$

is possible, under which the required steps is $\mathcal{O}(n + \sqrt{nL/\mu})\log(1/\epsilon)$.

Experiments

Experiments

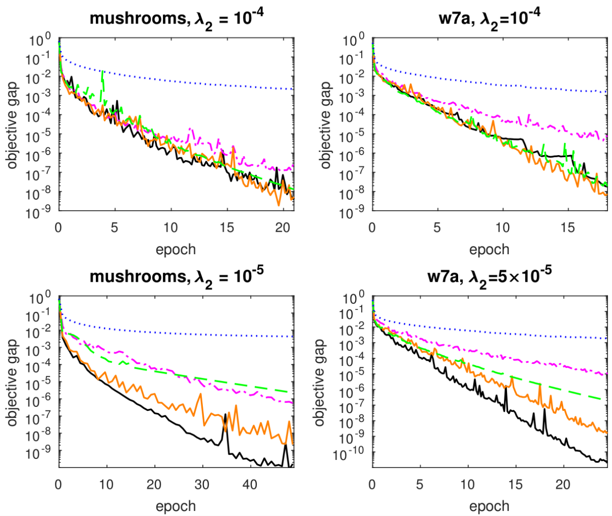

Figure 2: Comparison of several algorithms with $\ell_1\ell_2$-Logistic Regression.

Experiments

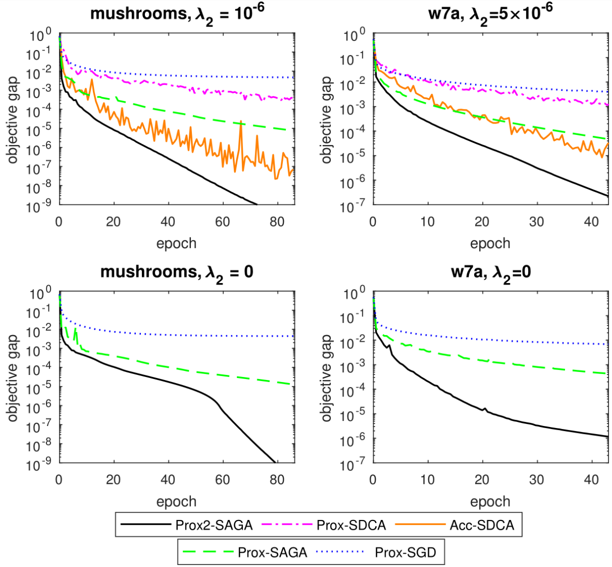

Figure 3: Comparison of several algorithms with $\ell_1\ell_2$-Logistic Regression.

Experiments

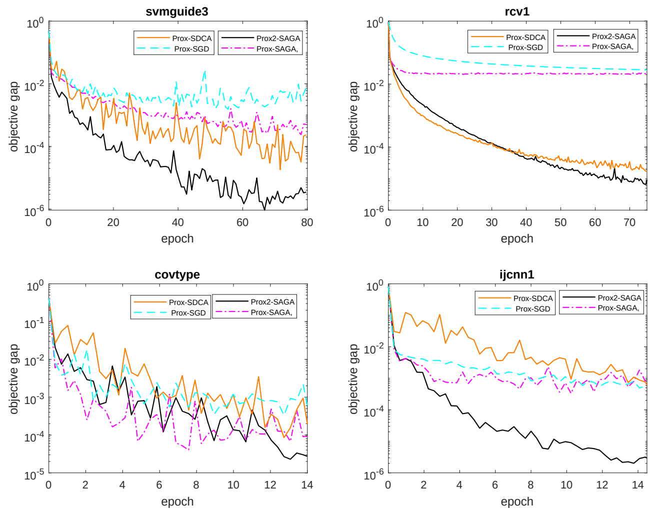

Figure 4: Comparison of several algorithms with sparse SVMs.

Conclusions

- Prox2-SAGA has combined Point-SAGA and DR splitting.
- Point-SAGA provides faster convergence rate to Prox2-SAGA.
- DR splitting provides the effectiveness.

Q & A