An Accelerated Variance Reducing Stochastic Method with Douglas-Rachford Splitting

Jingchang Liu[1], Linli Xu[1], Shuheng Shen[1], Qing Ling[2]

[1]School of Computer Science and Technology, University of Science and Technology

[2]School of Data and Computer Science, Sun Yat-Sen University

Background

Formulation

- Regularized ERM: $\min_{x\in\mathbb{R}^d} f(x)+h(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)+h(x)$.
- $f_i:\mathbb{R}^d\to\mathbb{R}$: 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:
  $\mathrm{prox}_f^{\gamma}(x) = \mathrm{argmin}_{y\in\mathbb{R}^d}\left(f(y)+\frac{1}{2\gamma}\|y-x\|^2\right)$.
- Gradient mapping: $f(x) = \frac{1}{\gamma}(x-\mathrm{prox}_f^{\gamma}(x))$.
- Subdifferential: $\partial f(x) = \{g \mid g^T(y-x)\le f(y)-f(x), \forall y\in \mathrm{dom}\, f\}$.
- Strongly convex: $f(y)\ge f(x)+\langle g, y-x\rangle+\frac{\mu}{2}\|y-x\|^2$.
- L-smooth: $f(y)\le f(x)+\langle\nabla f(x), y-x\rangle+\frac{L}{2}\|y-x\|^2$.

Exsiting Algorithm: $\mathrm{prox}_h^{\gamma}(x-\gamma\cdot\square)$, 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.

Convergence Rate

- VR stochastic methods: $\mathcal{O}((n+L/\mu)\log(1/\epsilon))$.
- 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: $f^{\gamma}(x)=\inf_y\left\{f(y)+\frac{1}{2\gamma}\|x-y\|^2\right\}$.

- 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 better than L/μ of f.
- Application: Point-SAGA, which is used when h is absent. At step $k+1$:
  $$z_j^k = x^k+\gamma(g_j^k-\sum_{i=1}^{n} g_i^k/n),$$
  $$x^{k+1}=\mathrm{prox}_{f_j}^{\gamma}(z_j^k)$$
  $$g_j^{k+1}=(z_j^k-x^{k+1})/\gamma,$$
  which is equivalent to $x^{k+1}=x^k-\gamma(g_j^{k+1}-g_j^k+\sum_{i=1}^{n} g_i^k/n)$, where g_j^{k+1} is the gradient mapping of f at z_j^k.

DR Splitting

- Formulation: $\min_x f(x)+h(x)$.
- Aim: Splitting the proximal operators of f and h.
- Iteration:
  $$y^{k+1} = -x^k+y^k+\mathrm{prox}_f^{\gamma}(2x^k-y^k),$$
  $$x^{k+1}=\mathrm{prox}_h^{\gamma}(y^{k+1}).$$

Methods

The Proposed Algorithm

Algorithm 1 Prox2-SAGA

1: **Input:** $x^0\in\mathbb{R}^d$, g_i^0 $(i=1,2,\ldots,n)$, step size $\gamma>0$.
2: **for** $k=0,1,\ldots$ **do**
3: Uniformly randomly pick j from 1 to n.
4: Calculate g_j^{k+1}:
$$z_j^k = x^k+\gamma\Big(g_j^k-\frac{1}{n}\sum_{i=1}^{n} g_i^k\Big),$$
$$g_j^{k+1}=\frac{1}{\gamma}\big((z_j^k+x^k-y^k)-\mathrm{prox}_{f_j}^{\gamma}(z_j^k+x^k-y^k)\big).$$
5: Update x:
$$y^{k+1}=z_j^k-\gamma g_j^{k+1},$$
$$x^{k+1}=\mathrm{prox}_h^{\gamma}(y^{k+1}).$$
6: Update g_i $(i=1,2,\ldots,n)$ in the table:
$$g_i^{k+1}=\begin{cases} g_j^{k+1}, & \text{if } i=j,\\ g_i^k, & \text{otherwise.}\end{cases}$$
7: **end for**
8: **Output:** x^{k+1}.

Figure 1: Prox2-SAGA Algorithm

Remarks

- Prox2-SAGA is under the algorithm framework of SAGA; it also involves two proximal operators. So it gets its name.
- When $h=0$, Prox2-SAGA is simplified to Point-SAGA.
- When $n=1$, Prox2-SAGA is simplified to DR-Splitting.

Main Iterations:
$$y^{k+1}=x^k-\gamma\Big(g_j^{k+1}-g_j^k+\sum_{i=1}^{n} g_i^k/n\Big),\quad x^{k+1}=\mathrm{prox}_h^{\gamma}(y^{k+1}),$$
where g_j^{k+1}, which is stored in a table, is the gradient mapping of f at $z_j^k+x^k-y^k$.

Main Theories

- **Proposition:** Suppose that $(y^{\infty},\{g_i^{\infty}\}_{i=1,\ldots,n})$ is the fixed point of the Prox2-SAGA iteration. Then $x^{\infty}=\mathrm{prox}_h^{\gamma}(y^{\infty})$ is a minimizer of the proposed problem.
- **Non-strongly convex case:** 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 $\gamma\le 1/L$, at any time $k>0$ it holds
  $$\mathbb{E}\|\bar{g}_j^k-g_j^*\|^2\le\frac{1}{k}\Big(\sum_{i=1}^{n}\|g_i^0-g_i^*\|^2+\|\frac{1}{\gamma}(y^0-y^*)\|^2\Big).$$
- **Strongly convex case:** 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
  $$\mathbb{E}\|x^k-x^*\|^2\le\Big(1-\frac{\mu\gamma}{2\mu\gamma+2}\Big)^k\cdot\frac{\mu\gamma-2}{2-n\mu\gamma}\Big\{\sum_{i=1}^{n}\|\gamma(g_i^0-g_i^*)\|^2+\|y^0-y^*\|^2\Big\}.$$
- **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 $\mathbb{E}\|x^k-x^*\|^2\le\epsilon$.
  - 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

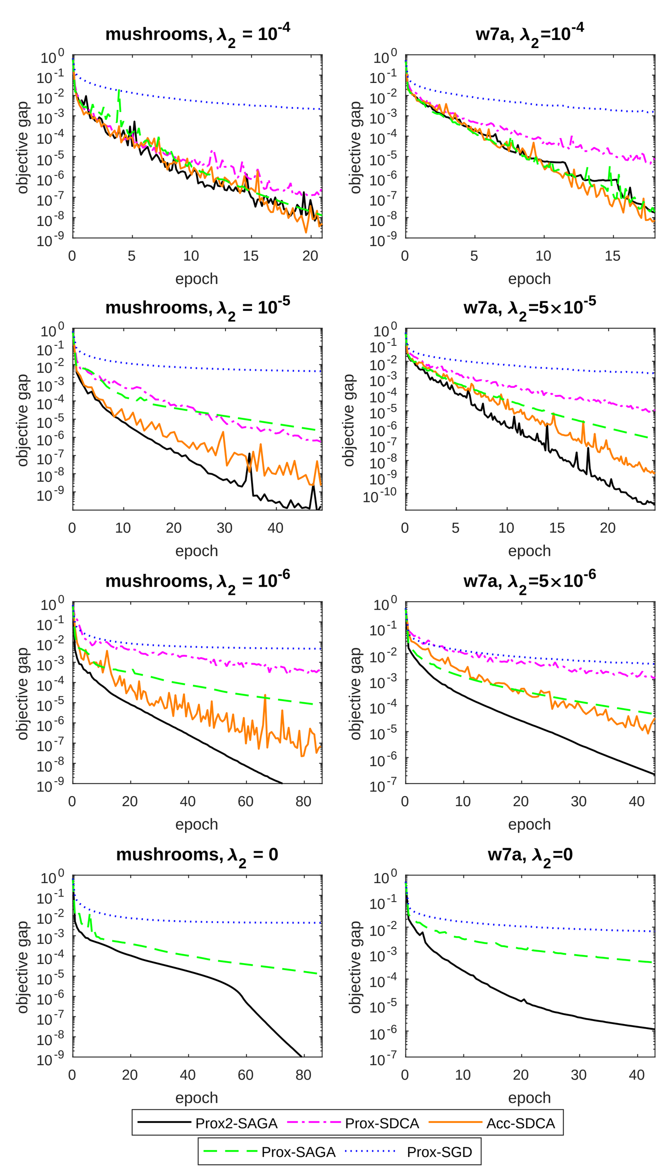

Figure 2: Comparison of several algorithms with $\ell_1\ell_2$-Logistic Regression

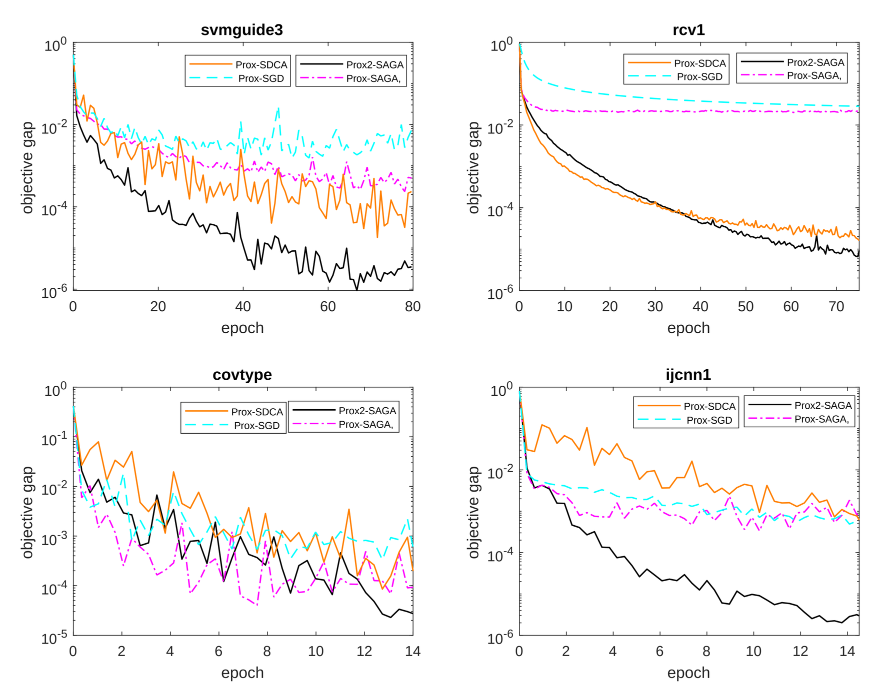

Figure 3: Comparison of several algorithms with sparse SVMs.

- Figure 2 shows that Prox2-SAGA has the accelerated effect when f_i's are ill-conditioned.
- Figure 3 shows that Prox2-SAGA would also work when f_i's are non-smooth.