ADMM for Stochastic Optimization

Hailiang ZHAO @ ZJU-CS

http://hliangzhao.me

October 30, 2022

Outline

I Stochastic ADMM
SADMM

II The Variation Reduction Technique
SVRG-ADMM

III Fusing VR with Momentum
Acc-SADMM

IV Stochastic Nonconvex Optimization
NC-SADMM
SPIDER-ADMM

References

Outline

I Stochastic ADMM
SADMM

II The Variation Reduction Technique
SVRG-ADMM

III Fusing VR with Momentum
Acc-SADMM

IV Stochastic Nonconvex Optimization
NC-SADMM
SPIDER-ADMM

References

Stochastic Optimization

Consider the following linearly constrained separable optimization problem:

$$\min_{\boldsymbol{x}_1, \boldsymbol{x}_2} \quad f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2), \quad s.t. \quad \mathbf{A}_1 \boldsymbol{x}_1 + \mathbf{A}_2 \boldsymbol{x}_2 = \boldsymbol{b}. \tag{1}$$

We assume that

$$f_1(\boldsymbol{x}_1) \equiv \mathbb{E}_\xi [F(\boldsymbol{x}_1; \xi)], \tag{2}$$

where $F(\boldsymbol{x}_1; \xi)$ is a stochastic component indexed by a random number ξ. For traditional machine learning, the data are often finitely sampled. If we denote each component function as $F_i(\boldsymbol{x})$, we can rewrite $f_1(\boldsymbol{x}_1)$ as below:

$$f_1(\boldsymbol{x}_1) \equiv \frac{1}{n} \sum_{i=1}^{n} F_i(\boldsymbol{x}_1). \tag{3}$$

Stochastic Optimization

When n is finite, (3) is an offline problem, with examples including empirical risk minimization. n can also go to infinity, which is a general case. In the following, when we study the finite-sum (offline) problem, we shall use the formula (3); otherwise, we use (2).

When n is large, accessing the exact function value of $f_1(\boldsymbol{x}_1)$ or its gradient may be very expensive and even impossible when $n = \infty$. To deal with such large-scale problems, the standard way is to *estimate the full gradient via one or several randomly sampled counterparts from individual functions*. We call algorithms using this technique as *stochastic algorithms*.

Stochastic ADMM

We consider (1) with (2).

In each iteration, we independently sample a stochastic index ξ and compute the stochastic gradient $\nabla F(\boldsymbol{x}_1, \xi)$ (denote by $\tilde{\nabla} f_1(\boldsymbol{x}_1)$).

We firstly give *SADMM*:

1. $\boldsymbol{x}_1^{k+1} = \mathrm{argmin}_{\boldsymbol{x}_1} \hat{L}_\beta^k(\boldsymbol{x}_1, \boldsymbol{x}_2^k, \boldsymbol{\lambda}^k)$
2. $\boldsymbol{x}_2^{k+1} = \mathrm{argmin}_{\boldsymbol{x}_2} \hat{L}_\beta^k(\boldsymbol{x}_1^{k+1}, \boldsymbol{x}_2, \boldsymbol{\lambda}^k)$
3. $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \beta(\mathbf{A}_1 \boldsymbol{x}_1^{k+1} + \mathbf{A}_2 \boldsymbol{x}_2^{k+1} - \boldsymbol{b})$

Wherein, the approximated augmented term is:

$$\hat{L}_\beta(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{\lambda}) = f_1(\boldsymbol{x}_1^k) + \langle \tilde{\nabla} f_1(\boldsymbol{x}_1^k), \boldsymbol{x}_1 - \boldsymbol{x}_1^k \rangle + f_2(\boldsymbol{x}_2) + \frac{\beta}{2}\|\mathbf{A}_1 \boldsymbol{x}_1 + \mathbf{A}_2 \boldsymbol{x}_2 - \boldsymbol{b} + \frac{1}{\beta}\boldsymbol{\lambda}\|^2 + \frac{1}{2\eta_{k+1}}\|\boldsymbol{x}_1 - \boldsymbol{x}_1^k\|^2. \quad (4)$$

Convergence Analysis for SADMM

Lemma I.1

Assume that f_1 is μ-strongly convex and L-smooth and f_2 is convex. For $k \geq 0$, if the step size $\eta_{k+1} \leq 1/(2L)$, then for any $\tilde{\boldsymbol{\lambda}}$, we have

$$\begin{aligned}
& f_1(\boldsymbol{x}_1^{k+1}) + f_2(\boldsymbol{x}_2^{k+1}) - f_1(\boldsymbol{x}_1^*) - f_2(\boldsymbol{x}_2^*) + \\
& \qquad\qquad \langle \tilde{\boldsymbol{\lambda}}, \mathbf{A}_1 \boldsymbol{x}_1^{k+1} + \mathbf{A}_2 \boldsymbol{x}_2^{k+1} - \boldsymbol{b} \rangle \\
& \leq \eta_{k+1} \|\tilde{\nabla} f_1(\boldsymbol{x}_1^k) - \nabla f_1(\boldsymbol{x}_1^k)\|^2 + (\frac{1}{2\eta_{k+1}} - \frac{\mu}{2}) \|\boldsymbol{x}_1^k - \boldsymbol{x}_1^*\|^2 \\
& - \frac{1}{2\eta_{k+1}} \|\boldsymbol{x}_1^{k+1} - \boldsymbol{x}_1^*\|^2 + \langle \nabla f_1(\boldsymbol{x}_1^k) - \tilde{\nabla} f_1(\boldsymbol{x}_1^k), \boldsymbol{x}_1^k - \boldsymbol{x}_1^* \rangle \\
& + \frac{1}{2\beta} (\|\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^k\|^2 - \|\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^{k+1}\|^2) \\
& + \frac{\beta}{2} (\|\mathbf{A}_2 \boldsymbol{x}_2^k - \mathbf{A}_2 \boldsymbol{x}_2^*\|^2 - \|\mathbf{A}_2 \boldsymbol{x}_2^{k+1} - \mathbf{A}_2 \boldsymbol{x}_2^*\|^2). \qquad (5)
\end{aligned}$$

Convergence Analysis for SADMM

Theorem I.1

Under the assumption of Lemma I.1, assume that the variance of f_1's gradient is uniformly bounded by σ^2, i.e.,

$$\mathbb{E}_\xi\left[\|\nabla F_1(\boldsymbol{x}_1, \boldsymbol{\xi}) - \nabla f_1(\boldsymbol{x}_1)\|^2\right] \leq \sigma^2, \forall \boldsymbol{x}_1. \tag{6}$$

Define

$$D_1 = \|\boldsymbol{x}_1^0 - \boldsymbol{x}_1^*\| \text{ and } D_2 = \|\mathbf{A}_2\boldsymbol{x}_2^0 - \mathbf{A}_2\boldsymbol{x}_2^*\|. \tag{7}$$

For the generally convex case, i.e., $\mu = 0$, set the step size $\eta_k = \frac{1}{2L+\sqrt{k}\sigma/D_1}$,

$$\bar{\boldsymbol{x}}_1^K = \frac{1}{\sum_{k=1}^K \eta_k} \sum_{k=1}^K \eta_k \boldsymbol{x}_1^k \text{ and } \bar{\boldsymbol{x}}_2^K = \frac{1}{\sum_{k=1}^K \eta_k} \sum_{k=1}^K \eta_k \boldsymbol{x}_2^k. \tag{8}$$

Convergence Analysis for SADMM

Theorem I.1 (cont'd)

Then, for any $\rho > 0$ and sufficiently large K, we have

$$\begin{aligned}\mathbb{E}[f_1(\bar{\boldsymbol{x}}_1^K)] + \mathbb{E}[f_2(\bar{\boldsymbol{x}}_2^K)] - f_1(\boldsymbol{x}_1^*) - f_2(\boldsymbol{x}_2^*) \\ + \rho\mathbb{E}\left[\|\mathbf{A}_1\bar{\boldsymbol{x}}_1^K + \mathbf{A}_2\bar{\boldsymbol{x}}_2^K - \boldsymbol{b}\|\right] \\ \leq \frac{2D_1\sigma\log K}{\sqrt{K}} + \frac{\sigma}{\sqrt{K}}\left[\frac{D_1}{2} + \frac{\rho^2}{2\beta(2LD_1+\sigma)} + \frac{\beta D_2^2}{2(2LD_1+\sigma)}\right].\end{aligned} \tag{9}$$

Convergence Analysis for SADMM

Theorem I.1 (cont'd)

For the strongly convex case, i.e., $\mu > 0$, set the step size $\eta_k = \frac{1}{2L+k\mu}$,

$$\bar{\boldsymbol{x}}_1^K = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_1^k \text{ and } \bar{\boldsymbol{x}}_2^K = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_2^k. \tag{10}$$

Then for any $\rho > 0$ we have

$$\begin{aligned} &\mathbb{E}[f_1(\bar{\boldsymbol{x}}_1^K)] + \mathbb{E}[f_2(\bar{\boldsymbol{x}}_2^K)] - f_1(\boldsymbol{x}_1^*) - f_2(\boldsymbol{x}_2^*) \\ &\qquad\qquad + \rho\mathbb{E}\left[\|\mathbf{A}_1\bar{\boldsymbol{x}}_1^K + \mathbf{A}_2\bar{\boldsymbol{x}}_2^K - \boldsymbol{b}\|\right] \\ &\leq \frac{\sigma^2(\log K + 1)}{\mu K} + \frac{1}{K}\left[LD_1^2 + \frac{\rho^2}{2\beta} + \frac{\beta D_2^2}{2}\right]. \end{aligned} \tag{11}$$

Outline

I Stochastic ADMM
SADMM

II The Variation Reduction Technique
SVRG-ADMM

III Fusing VR with Momentum
Acc-SADMM

IV Stochastic Nonconvex Optimization
NC-SADMM
SPIDER-ADMM

References

Variance Reduction

The Variance Reduction (VR) technique is initially designed to solve the problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} F_i(\boldsymbol{x}). \tag{12}$$

It is known that the standard Stochastic Gradient Descent (SGD) will enjoy a sublinear convergence rate when each F_i is strongly convex and smooth. Surprisingly, the VR technique can accelerate stochastic algorithms to a linear convergence rate. The VR method uses the sum of the latest individual gradients as an estimator. The method requires $\mathcal{O}(nd)$ memory storage and the estimated gradient is a biased gradient estimator.

Variance Reduction

In the following, we introduce the application of VR to ADMM methods. We show that for the offline problems, VR improves the convergence rate to $\mathcal{O}(1/K)$ for the generally convex case. We use a classical VR method called SVRG. Its main technique is to *frequently pre-store a snapshot vector and to control the variance via the snapshot vector and the latest iterate.*

Specifically, we consider (1) with (3). In the process of solving the primal variable, we linearize both $f_1(\boldsymbol{x}_1)$ and the augmented term $\frac{\beta}{2}\|\mathbf{A}_1\boldsymbol{x}_1 + \mathbf{A}_2\boldsymbol{x}_2 - \boldsymbol{b} + \frac{1}{\beta}\boldsymbol{\lambda}\|^2$.

SVRG-ADMM

We have the following intermediate iterations:

$$\begin{aligned}
\boldsymbol{x}_{s,1}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}_1} \Big(&\langle \tilde{\nabla} f_1(\boldsymbol{x}_{s,1}^k), \boldsymbol{x}_1 - \boldsymbol{x}_{s,1}^k \rangle \\
&+ \langle \beta(\mathbf{A}_1 \boldsymbol{x}_{s,1}^k + \mathbf{A}_2 \boldsymbol{x}_{s,2}^k - \boldsymbol{b}) + \boldsymbol{\lambda}_s^k, \mathbf{A}_1(\boldsymbol{x}_1 - \boldsymbol{x}_{s,1}^k) \rangle \\
&+ \frac{1}{2\eta_1} \|\boldsymbol{x}_1 - \boldsymbol{x}_{s,1}^k\|^2 \Big), \qquad (13) \\
\boldsymbol{x}_{s,2}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}_2} \Big(&f_2(\boldsymbol{x}_2) \\
&+ \langle \beta(\mathbf{A}_1 \boldsymbol{x}_{s,1}^{k+1} + \mathbf{A}_2 \boldsymbol{x}_{s,2}^k - \boldsymbol{b}) + \boldsymbol{\lambda}_s^k, \mathbf{A}_2(\boldsymbol{x}_2 - \boldsymbol{x}_{s,2}^k) \rangle \\
&+ \frac{1}{2\eta_2} \|\boldsymbol{x}_2 - \boldsymbol{x}_{s,2}^k\|^2 \Big), \qquad (14)
\end{aligned}$$

where

$$\eta_1 = \frac{1}{9L + \beta\|\mathbf{A}_1\|^2} \text{ and } \eta_2 = \frac{1}{\beta\|\mathbf{A}_2\|^2}. \qquad (15)$$

SVRG-ADMM

We noe presente the detailed procedures. We call it *SVRG-ADMM*.

1. **for** $s = 0, ..., S-1$ **do**
   - 1.1 **for** $k = 0, ..., m-1$ **do**
     - 1.1.1 Randomly sample $i_{k,s}$ from $[n]$
     - 1.1.2 $\tilde{\nabla} f_1(\boldsymbol{x}_{s,1}^k) = \nabla F_{i_{k,s}}(\boldsymbol{x}_{s,1}^k) - \nabla F_{i_{k,s}}(\tilde{\boldsymbol{x}}_{s,1}) + \frac{1}{n}\sum_{i=1}^n \nabla F_i(\tilde{\boldsymbol{x}}_{s,1})$
     - 1.1.3 Update $\boldsymbol{x}_{s,1}^{k+1}$ and $\boldsymbol{x}_{s,2}^{k+1}$ by (13) and (14), respectively
     - 1.1.4 $\boldsymbol{\lambda}_s^{k+1} = \boldsymbol{\lambda}_s^k + \beta(\mathbf{A}_1\boldsymbol{x}_{s,1}^{k+1} + \mathbf{A}_2\boldsymbol{x}_{s,2}^{k+1} - \boldsymbol{b})$
   - 1.2 $\tilde{\boldsymbol{x}}_{s+1,i} = \frac{1}{m}\sum_{k=1}^m \boldsymbol{x}_{s,i}^k$
   - 1.3 $\boldsymbol{x}_{s+1,i}^0 = \boldsymbol{x}_{s,i}^m$, $i = 1, 2$
   - 1.4 $\boldsymbol{\lambda}_{s+1}^0 = \boldsymbol{\lambda}_s^m$

Step 1.1.2 is used to reduce the variance, in which $\tilde{\boldsymbol{x}}_{s,1}$ is the snapshot vector and $\frac{1}{n}\sum_{i=1}^n \nabla F_i(\tilde{\boldsymbol{x}}_{s,1})$ is re-computed at the beginning of the outer loop.

Convergence Analysis of SVRG-ADMM

In the following, we show that the variance of this estimated gradient can be controlled.

Lemma II.1

Assume that F_i is convex and L-smooth for all $i \in [n]$. Let $\mathbb{E}_k$ denote the expectation taken only on the random number $i_{k,s}$ conditioned on $\boldsymbol{x}_{s,1}^k$. Then we have

$$\mathbb{E}_k[\tilde{\nabla} f_1(\boldsymbol{x}_{s,1}^k)] = \nabla f_1(\boldsymbol{x}_{s,1}^k). \tag{16}$$

We have

$$\mathbb{E}_k\Big[\|\tilde{\nabla} f_1(\boldsymbol{x}_{s,1}k) - \nabla f_1(\boldsymbol{x}_{s,1}^k)\|^2\Big] \leq 4L\Big[H_1(\boldsymbol{x}_{s,1}^k) + H_1(\tilde{\boldsymbol{x}}_{s,1})\Big], \tag{17}$$

where $H_1(\boldsymbol{x}_1) = f_1(\boldsymbol{x}_1) + f_1(\boldsymbol{x}_1^*) - \langle \nabla f_1(\boldsymbol{x}_1^*), \boldsymbol{x}_1 - \boldsymbol{x}_1^* \rangle$.

Convergence Analysis of SVRG-ADMM

In the following, we study the inner loop. For the sake of simplicity, we drop the subscript s in the analysis of inner loop, since it is clear from the context.

Lemma II.2

Assume that F_i is convex and L-smooth for $i \in [n]$ and f_2 is convex. Then for $k \geq 0$,

$$\begin{aligned}
&\mathbb{E}_k[f_1(\boldsymbol{x}_1^{k+1})] - f_1(\boldsymbol{x}_1^*) + \mathbb{E}_k[f_2(\boldsymbol{x}_2^{k+1})] - f_2(\boldsymbol{x}_2^*) \\
&\qquad + \mathbb{E}_k\big[\langle \boldsymbol{\lambda}^*, \mathbf{A}_1\boldsymbol{x}_1^{k+1} + \mathbf{A}_2\boldsymbol{x}_2^{k+1} - \boldsymbol{b}\rangle\big] \\
&\leq \frac{1}{4}\Big(H_1(\boldsymbol{x}_1^k) + H_1(\tilde{\boldsymbol{x}}_1)\Big) + \|\boldsymbol{x}_1^k - \boldsymbol{x}_1^*\|_{\mathbf{G}_1}^2 - \mathbb{E}_k[\|\boldsymbol{x}_1^{k+1} - \boldsymbol{x}_1^*\|_{\mathbf{G}_1}^2] \\
&\qquad + \|\boldsymbol{x}_2^k - \boldsymbol{x}_2^*\|_{\mathbf{G}_2}^2 - \mathbb{E}_k[\|\boldsymbol{x}_2^{k+1} - \boldsymbol{x}_2^*\|_{\mathbf{G}_2}^2] \\
&\qquad + \frac{1}{2\beta}\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^k\|^2 - \frac{1}{2\beta}\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{k+1}\|^2,
\end{aligned} \tag{18}$$

where $\mathbf{G}_1 = \frac{1}{2}[(\beta\|\mathbf{A}_1\|^2 + 9L)\mathbf{I} - \beta\mathbf{A}_1^T\mathbf{A}_1]$, and $\mathbf{G}_2 = \frac{\beta}{2}\|\mathbf{A}_2\|^2\mathbf{I}$.

Convergence Analysis of SVRG-ADMM

Theorem II.1

Under the assumption of Lemma II.2, letting

$$D_\lambda = \|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}_0^0\|, \tag{19}$$

$$D_i = \|\boldsymbol{x}_{0,i}^0 - \boldsymbol{x}_i^*\|_{\mathbf{G}_i}, \, i = 1, 2, \tag{20}$$

$$D_f = f_1(\boldsymbol{x}_{0,1}^0) - f_1(\boldsymbol{x}_1^*) - \langle \nabla f_1(\boldsymbol{x}_1^*), \boldsymbol{x}_{0,1}^0 - \boldsymbol{x}_1^* \rangle, \tag{21}$$

$$\bar{\boldsymbol{x}}_i^S = \frac{1}{S} \sum_{s=1}^{S} \tilde{\boldsymbol{x}}_{s,i}, \, i = 1, 2. \tag{22}$$

Convergence Analysis of SVRG-ADMM

Theorem II.1 (cont'd)

Then for SVRG-ADMM, we have

$$\mathbb{E}\Big[f_1(\bar{\boldsymbol{x}}_1^S) + f_2(\bar{\boldsymbol{x}}_2^S) - f_1(\boldsymbol{x}_1^*) - f_2(\boldsymbol{x}_2^*) + \langle \boldsymbol{\lambda}^*, \mathbf{A}_1\bar{\boldsymbol{x}}_1^S + \mathbf{A}_2\bar{\boldsymbol{x}}_2^S - \boldsymbol{b}\rangle\Big] \le \frac{(m+1)D_f}{2Sm} + \frac{D_\lambda^2}{\beta mS} + \frac{2(D_1^2 + D_2^2)}{mS}, \tag{23}$$

$$\mathbb{E}\big[\|\mathbf{A}_1\bar{\boldsymbol{x}}_1^S + \mathbf{A}_2\bar{\boldsymbol{x}}_2^S - \boldsymbol{b}\|\big] \le \frac{D_\lambda}{m\beta S} + \frac{\sqrt{D_\lambda^2 + 2\beta(D_1^2 + D_2^2) + \frac{\beta(m+1)}{2}D_f}}{m\beta S}. \tag{24}$$

For generic nonconvex optimization, SVRG-ADMM can only achieve a complexity of $\mathcal{O}(\min(\epsilon^{-10/3}, n + n^{2/3}\epsilon^{-2}))$.

Outline

I Stochastic ADMM
SADMM

II The Variation Reduction Technique
SVRG-ADMM

III Fusing VR with Momentum
Acc-SADMM

IV Stochastic Nonconvex Optimization
NC-SADMM
SPIDER-ADMM

References

Momentum Acceleration

When applying the VR technique, the algorithms are transformed to act like a deterministic algorithm. So it is possible to fuse the momentum technique.

We consider the convex finite-sum problem with linear constraints in the general setting:

$$\begin{aligned} \min_{\boldsymbol{x}_1, \boldsymbol{x}_2} & \left(h_1(\boldsymbol{x}_1) + f_1(\boldsymbol{x}_1) + h_2(\boldsymbol{x}_2) + \frac{1}{n} \sum_{i=1}^{n} F_{2,i}(\boldsymbol{x}_2) \right), \\ s.t. \quad & \mathbf{A}_1 \boldsymbol{x}_1 + \mathbf{A}_2 \boldsymbol{x}_2 = \boldsymbol{b}, \end{aligned} \tag{25}$$

where $f_1(\boldsymbol{x}_1)$ and $F_{2,i}(\boldsymbol{x}_2)$ with $i \in [n]$ are convex and L_1-smooth and L_2-smooth, respectively, and $h_1(\boldsymbol{x}_1)$ and $h_2(\boldsymbol{x}_2)$ are also convex and their proximal mappings can be solved efficiently.

Acc-SADMM

We define

$$f_2(\boldsymbol{x}_2) = \frac{1}{n}\sum_{i=1}^{n} F_{2,i}(\boldsymbol{x}_2),$$
$$\mathcal{J}_1(\boldsymbol{x}_1) = h_1(\boldsymbol{x}_1) + f_1(\boldsymbol{x}_1), \mathcal{J}_2(\boldsymbol{x}_2) = h_2(\boldsymbol{x}_2) + f_2(\boldsymbol{x}_2),$$
$$\boldsymbol{x} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T, \mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2], \mathcal{J}(\boldsymbol{x}) = \mathcal{J}_1(\boldsymbol{x}_1) + \mathcal{J}_2(\boldsymbol{x}_2).$$

The method we will introduce is called *Acc-SADMM*. It also has double loops.

The update of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are as follows.

$$\begin{aligned}\boldsymbol{x}_{s,1}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}_1} \Big[& h_1(\boldsymbol{x}_1) + \langle \nabla f_1(\boldsymbol{y}_{s,1}^k), \boldsymbol{x}_1 \rangle \\ & + \langle \frac{\beta}{\theta_{1,s}}(\mathbf{A}_1 \boldsymbol{y}_{s,1}^k + \mathbf{A}_2 \boldsymbol{y}_{s,2}^k - \boldsymbol{b}) + \boldsymbol{\lambda}_s^k, \mathbf{A}_1 \boldsymbol{x}_1 \rangle \\ & + \left(\frac{L_1}{2} + \frac{\beta}{2\theta_{1,s}} \|\mathbf{A}_1\|^2 \right) \|\boldsymbol{x}_1 - \boldsymbol{y}_{s,1}^k\|^2 \Big]. \end{aligned} \tag{26}$$

Acc-SADMM

The update of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are as follows (cont'd).

$$
\begin{aligned}
\boldsymbol{x}_{s,2}^{k+1} = \operatorname*{argmin}_{\boldsymbol{x}_2} \Big\{ & h_2(\boldsymbol{x}_2) + \langle \tilde{\nabla} f_2(\boldsymbol{y}_{s,2}^k), \boldsymbol{x}_2 \rangle \\
& + \langle \frac{\beta}{\theta_{1,s}} (\mathbf{A}_1 \boldsymbol{x}_{s,1}^{k+1} + \mathbf{A}_2 \boldsymbol{y}_{s,2}^k - \boldsymbol{b}) + \boldsymbol{\lambda}_s^k, \mathbf{A}_2 \boldsymbol{x}_2 \rangle \\
& + \Big[\frac{1}{2}(1 + \frac{1}{b\theta_2}) L_2 + \frac{\beta}{2\theta_{1,s}} \|\mathbf{A}_2\|^2 \Big] \|\boldsymbol{x}_2 - \boldsymbol{y}_{s,2}^k\|^2 \Big\},
\end{aligned} \tag{27}
$$

where

$$
\tilde{\nabla} f_2(\boldsymbol{y}_{s,2}^k) = \frac{1}{b} \sum_{i_{k,s} \in \mathcal{I}_{k,s}} \Big(\nabla F_{2,i_{k,s}}(\boldsymbol{y}_{s,2}^k) - \nabla F_{2,i_{k,s}}(\tilde{\boldsymbol{x}}_{s,2}) + \nabla f_2(\tilde{\boldsymbol{x}}_{s,2}) \Big), \tag{28}
$$

whre $\mathcal{I}_{k,s}$ is a mini-batch of indices randomly drawn from $[n]$ with a size of b.

Acc-SADMM

With the above iterations, the inner loop of Acc-SADMM is as follows. Note that $\boldsymbol{y}_{s,1}^k$ and $\boldsymbol{y}_{s,2}^k$ are extrapolation variables.

1. **for** $k = 0, \ldots, m-1$ **do**
   1.1 $\boldsymbol{\lambda}_s^k = \tilde{\boldsymbol{\lambda}}_s^k + \frac{\beta\theta_2}{\theta_{1,s}}(\mathbf{A}_1\boldsymbol{x}_{s,1}^k + \mathbf{A}_2\boldsymbol{x}_{s,2}^k - \tilde{\boldsymbol{b}}_s)$
   1.2 Update $\boldsymbol{x}_{s,1}^{k+1}$ by (26)
   1.3 Update $\boldsymbol{x}_{s,2}^{k+1}$ by (27)
   1.4 $\tilde{\boldsymbol{\lambda}}_s^{k+1} = \boldsymbol{\lambda}_s^k + \beta(\mathbf{A}_1\boldsymbol{x}_{s,1}^{k+1} + \mathbf{A}_2\boldsymbol{x}_{s,2}^{k+1} - \boldsymbol{b})$
   1.5 $\boldsymbol{y}_s^{k+1} = \boldsymbol{x}_s^{k+1} + (1 - \theta_{1,s} - \theta_2)(\boldsymbol{x}_s^{k+1} - \boldsymbol{x}_s^k)$

Acc-SADMM

Acc-SADMM is demonstrated as follows.

1. Initialize parameters
2. **for** $s = 0, \dots, S-1$ **do**
   2.1 Do innner loop as the previous page says
   2.2 Set primal variables $\boldsymbol{x}^0_{s+1} = \boldsymbol{x}^m_s$
   2.3 $\tilde{\boldsymbol{x}}_{s+1} = \frac{1}{m}\left(\left[1 - \frac{(\tau-1)\theta_{1,s+1}}{\theta_2}\right]\boldsymbol{x}^m_s + \left[1 + \frac{(\tau-1)\theta_{1,s+1}}{(m-1)\theta_2}\right]\sum_{k=1}^{m-1}\boldsymbol{x}^k_s\right)$
   2.4 $\tilde{\boldsymbol{\lambda}}^0_{s+1} = \boldsymbol{\lambda}^{m-1}_s + \beta(1-\tau)(\mathbf{A}_1\boldsymbol{x}^m_{s,1} + \mathbf{A}_2\boldsymbol{x}^m_{s,2} - \boldsymbol{b})$
   2.5 $\boldsymbol{y}^0_{s+1} =$
   $(1-\theta_2)\boldsymbol{x}^m_s + \theta_2\tilde{\boldsymbol{x}}_{s+1} + \frac{\theta_{1,s+1}}{\theta_{1,s}}[(1-\theta_{1,s})\boldsymbol{x}^m_s - (1-\theta_{1,s}-\theta_2)\boldsymbol{x}^{m-1}_s - \theta_2\tilde{\boldsymbol{x}}_s]$
3. Output $\hat{\boldsymbol{x}}_S = \frac{1}{(m-1)(\theta_{1,S}+\theta_2)+1}\boldsymbol{x}^m_S + \frac{\theta_{1,S}+\theta_2}{(m-1)(\theta_{1,S}+\theta_2)+1}\sum_{k=1}^{m-1}\boldsymbol{x}^k_S$

The convergence rate of Acc-SADMM is $\mathcal{O}(1/S)$. The result is shown in Theorem 5.4 of the ADMM book (p190).

Outline

I Stochastic ADMM
SADMM

II The Variation Reduction Technique
SVRG-ADMM

III Fusing VR with Momentum
Acc-SADMM

IV Stochastic Nonconvex Optimization
NC-SADMM
SPIDER-ADMM

References

Nonconvex SADMM

We consider stochastic ADMM in the nonconvex setting. We study a two-block linearly constrained problem shown as follows:

$$\min_{\boldsymbol{x},\boldsymbol{y}} f(\boldsymbol{x}) + g(\boldsymbol{y}), \quad s.t. \quad \mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{y} = \boldsymbol{b}, \tag{29}$$

where $f(\boldsymbol{x}) = \mathbb{E}_\xi[F(\boldsymbol{x};\xi)]$.

Assumption IV.1

f and g are L_1-smooth and L_2-smooth, respectively. Moreover, the variance of stochastic gradients for f is uniformly bounded by σ^2, i.e.,

$$\mathbb{E}_\xi\left[\|\nabla F(\boldsymbol{x},\xi) - \nabla f(\boldsymbol{x})\|^2\right] \le \sigma^2, \forall \boldsymbol{x}. \tag{30}$$

NC-SADMM

We consider the following iterations:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \eta\left[\tilde{\nabla} f(\boldsymbol{x}^k) + \beta \mathbf{A}^T\left(\mathbf{A}\boldsymbol{x}^k + \mathbf{B}\boldsymbol{y}^k - \boldsymbol{b} + \frac{\boldsymbol{\lambda}^k}{\beta}\right)\right], \tag{31}$$

$$\boldsymbol{y}^{k+1} = \operatorname*{argmin}_{\boldsymbol{y}} \Big(g(\boldsymbol{y}) + \langle \boldsymbol{\lambda}^k, \mathbf{B}\boldsymbol{y}\rangle + \frac{\beta}{2}\|\mathbf{A}\boldsymbol{x}^{k+1} + \mathbf{B}\boldsymbol{y} - \boldsymbol{b}\|^2 + D_\phi(\boldsymbol{y}, \boldsymbol{y}^k)\Big), \tag{32}$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \beta\left(\mathbf{A}\boldsymbol{x}^{k+1} + \mathbf{B}\boldsymbol{y}^{k+1} - \boldsymbol{b}\right), \tag{33}$$

where $\tilde{\nabla} f(\boldsymbol{x}^k)$ is a stochastic estimator:

$$\tilde{\nabla} f(\boldsymbol{x}^k) = \frac{1}{S}\sum_{\xi \in \mathcal{I}_k} \nabla F(\boldsymbol{x}^k, \xi). \tag{34}$$

We call (31), (32), and (33) *NC-SADMM*.

Because the indices in $\mathcal{I}_k$ are drawn independently, we have

$$\mathbb{E}_k[\tilde{\nabla} f(\boldsymbol{x}^k)] = \nabla f(\boldsymbol{x}^k) \tag{35}$$

$$\mathbb{E}_k\left[\|\tilde{\nabla} f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{x}^k)\|^2\right] \leq \frac{\sigma^2}{S}, \tag{36}$$

where the expectation is taken under the condition that the previous k iterates are known. Besides, noe that the augmented Lagrangian function is

$$\begin{aligned} L_\beta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \mathrm{g}(\boldsymbol{y}) + \langle \boldsymbol{\lambda}, \mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{y} - \boldsymbol{b} \rangle \\ + \frac{\beta}{2}\|\mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{y} - \boldsymbol{b}\|^2. \end{aligned} \tag{37}$$

We can find that L_β is $\tilde{L}_1$-smooth w.r.t. $\boldsymbol{x}$ and L_β is $\tilde{L}_2$-smooth w.r.t. $\boldsymbol{y}$, where $\tilde{L}_1 = L_1 + \beta\|\mathbf{A}\|^2$ and $\tilde{L}_2 = L_2 + \beta\|\mathbf{B}\|^2$.

Convergence Analysis of NC-SADMM

We now show that NC-SADMM can find an ϵ-approximation KKT point in $\mathcal{O}(\epsilon^{-4})$ stochastic accesses of gradient in expectation.

Theorem IV.1

Assume that Assumption IV.1 holds and there exists $\mu > 0$ such that $\|\mathbf{B}^T\boldsymbol{\lambda}\| \geq \mu\|\boldsymbol{\lambda}\|$ for all $\boldsymbol{\lambda}$ (surjectiveness of $\mathbf{B}$). Set

$$\eta \in [\Theta(\epsilon^2), 1/\tilde{L}_1] \text{ and } S = \eta \cdot \Theta(\epsilon^2) \in \mathbb{Z}^+. \tag{38}$$

Pick ϕ to be $\rho = \Theta(1)$-strongly convex and $L = \Theta(1)$-smooth, set $\beta \geq \frac{24(L_2^2+2L^2)}{\mu^2\rho} = \Theta(1)$, and define the Lyapunov function:

$$\Phi^k = L_\beta(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{\lambda}^k) + \frac{6L^2}{\mu^2\beta}\|\boldsymbol{y}^k - \boldsymbol{y}^{k-1}\|^2. \tag{39}$$

Then after running NC-SADMM by $K = \eta^{-1}\epsilon^{-2}$ iterations, we find an $\mathcal{O}(\epsilon)$-approximate KKT point in expectation.

Convergence Analysis of NC-SADMM

Theorem IV.1 (cont'd)

Specifically, letting $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{\lambda}})$ uniformly randomly taken from $\{\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{\lambda}^k\}_{k=1}^K$, defining $D = \Phi^0 - \min_{k\geq 0} \mathbb{E}[\Phi^k]$, and assuming that D is finite, we have

$$\tilde{\mathbb{E}}\left[\|\mathbf{A}\tilde{\boldsymbol{x}} + \mathbf{B}\tilde{\boldsymbol{y}} - \boldsymbol{b}\|^2\right] \leq \frac{1}{\beta}\left(\frac{D}{K} + \frac{\tilde{L}_1 \eta^2 \sigma^2}{2S}\right) = \mathcal{O}(\epsilon^2), \tag{40}$$

$$\tilde{\mathbb{E}}\left[\|\nabla g(\tilde{\boldsymbol{y}}) + \mathbf{B}^T \tilde{\boldsymbol{\lambda}}\|^2\right] \leq \frac{4L^2}{\rho}\left(\frac{D}{K} + \frac{\tilde{L}_1 \eta^2 \sigma^2}{2S}\right) = \mathcal{O}(\epsilon^2), \tag{41}$$

$$\begin{aligned}\tilde{\mathbb{E}}\left[\|\nabla f(\tilde{\boldsymbol{x}}) + \mathbf{A}^T \tilde{\boldsymbol{\lambda}}\|^2\right] \leq 2\frac{K+1}{K}&\left(2 + \eta\beta\|\mathbf{A}\|^2\right) \\ &\left(\frac{D}{\eta(K+1)} + \frac{\tilde{L}_1 \eta \sigma^2}{2S}\right) = \mathcal{O}(\epsilon^2),\end{aligned} \tag{42}$$

where $\tilde{\mathbb{E}}$ denotes taking expectation for all the randomness in NC-SADMM and the selection of $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{\lambda}})$.

SPIDER-ADMM

The Stochastic Path-Integrated Differential Estimator (SPIDER) technique is a radical VR method that is used to track quantities using reduced stochastic oracles.

For generic L-smooth stochastic nonconvex optimization, SPIDER can achieve the optimal $\mathcal{O}(\epsilon^{-3})$ expected complexity to find an ϵ-approximate first- order stationary point. This result is different from variance reduction methods in the convex case, as the latter can only accelerate the convergence rate for the finite-sum problems.

We also note that for the finite-sum problem with n individual functions, SPIDER can improve the complexity to $\mathcal{O}(\min\{n + n^{1/2}\epsilon^{-2}, \epsilon^{-3}\})$.

SPIDER-ADMM

In the following, we apply the SPIDER technique to accelerate the nonconvex SADMM algorithm. We consider a multi-block linearly constrained problem shown as below:

$$\min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\boldsymbol{y}} \sum_{i=1}^{m} f_i(\boldsymbol{x}_i) + g(\boldsymbol{y}), \quad s.t. \quad \sum_{i=1}^{m} \mathbf{A}_i \boldsymbol{x}_i + \mathbf{B}\boldsymbol{y} = \boldsymbol{b}, \tag{43}$$

where $f_i(\boldsymbol{x}_i) = \mathbb{E}_{\xi_i}[F_i(\boldsymbol{x}_i; \xi_i)]$ for $i \in [m]$.

Assumption IV.2

g is L_0-smooth. For each $i \in [m]$, $F_i(\boldsymbol{x}_i; \xi_i)$ is L_i-smooth w.r.t. $\boldsymbol{x}_i$ for all ξ_i. Moreover, the variance of stochastic gradients of f_i is uniformly bounded by σ^2, i.e.,

$$\mathbb{E}_{\xi_i}\left[\|\nabla F_i(\boldsymbol{x}_i, \xi_i) - \nabla f_i(\boldsymbol{x}_i)\|^2\right] \leq \sigma^2, \forall \boldsymbol{x}. \tag{44}$$

We further define $\boldsymbol{x} = [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_m^T]$, $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_m]$.

SPIDER-ADMM

We consider the following iterations:

$$\begin{aligned}
\boldsymbol{x}_i^{k+1} = \boldsymbol{x}_i^k - \eta \Big[\tilde{\nabla} f_i(\boldsymbol{x}_i^k) + \beta \mathbf{A}_i^T \Big(\sum_{j<i} \mathbf{A}_j \boldsymbol{x}_j^{k+1} \\
+ \sum_{j \geq i} \mathbf{A}_j \boldsymbol{x}_j^k + \mathbf{B} \boldsymbol{y}^k - \boldsymbol{b} + \frac{\boldsymbol{\lambda}^k}{\beta} \Big) \Big],
\end{aligned} \tag{45}$$

$$\begin{aligned}
\boldsymbol{y}^{k+1} = \operatorname*{argmin}_{\boldsymbol{y}} \Big(\mathrm{g}(\boldsymbol{y}) + \langle \boldsymbol{\lambda}^k, \mathbf{B} \boldsymbol{y} \rangle + \frac{\beta}{2} \| \mathbf{A} \boldsymbol{x}^{k+1} + \mathbf{B} \boldsymbol{y} - \boldsymbol{b} \|^2 \\
+ D_\phi(\boldsymbol{y}, \boldsymbol{y}^k) \Big),
\end{aligned} \tag{46}$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \beta \big(\mathbf{A} \boldsymbol{x}^{k+1} + \mathbf{B} \boldsymbol{y}^{k+1} - \boldsymbol{b} \big). \tag{47}$$

We call (45), (46), and (47) *SPIDER-ADMM*.

SPIDER-ADMM

In (45), $\tilde{\nabla} f_i(\boldsymbol{x}_i^k)$ is chosen as follows.

- For a certain hyper-parameter q, if the iteration k is divisible by q, then

$$\tilde{\nabla} f_i(\boldsymbol{x}_i^k) = \frac{1}{S_1} \sum_{\xi_i \in \mathcal{I}_{k,i}} \nabla F_i(\boldsymbol{x}_i^k, \xi_i), \tag{48}$$

  where $\mathcal{I}_{k,i}$ is a mini-batch of indices of size S_1.

- Otherwise,

$$\begin{aligned} \tilde{\nabla} f_i(\boldsymbol{x}_i^k) = \frac{1}{S_2} \sum_{\xi_i \in \mathcal{I}_{k,i}} \left[\nabla F_i(\boldsymbol{x}_i^k, \xi_i) - \nabla F_i(\boldsymbol{x}_i^{k-1}, \xi_i) \right] \\ + \tilde{\nabla} f_i(\boldsymbol{x}_i^{k-1}), \end{aligned} \tag{49}$$

  where $\mathcal{I}_{k,i}$ is a mini-batch of indices of size S_2.

Convergence Analysis of SPIDERS-ADMM

Similar to Theorem IV.1, we have the following convergence result.

Theorem IV.2

Assume that Assumption IV.2 holds and there exists $\mu > 0$ such that $\|\mathbf{B}^T\boldsymbol{\lambda}\| \geq \mu\|\boldsymbol{\lambda}\|$ for all $\boldsymbol{\lambda}$ (surjectiveness of $\mathbf{B}$). Set

$$S_1 = \Theta(\epsilon^{-2}), S_2 = \Theta(\epsilon^{-1}), q = \Theta(\epsilon^{-1}), \tag{50}$$

and

$$\eta = \min\left\{\frac{1}{2\max_{i\in[m]}\{\tilde{L}_i\}}, \frac{1}{2\max_{i\in[m]}\{L_i\}\sqrt{q/S_2}}\right\} = \Theta(1). \tag{51}$$

Note that $\tilde{L}_{=}L_i + \beta\|\mathbf{A}_i\|^2$, $\tilde{L}_0 = L_0 + \beta\|\mathbf{B}\|^2$.

Convergence Analysis of SPIDER-ADMM

Theorem IV.2 (cont'd)

Pick ϕ to be $\rho = \Theta(1)$-strongly convex and $L = \Theta(1)$-smooth, set $\beta \geq \frac{24(L_0^2+2L^2)}{\mu^2\rho} = \Theta(1)$, and define the Lyapunov function:

$$\Phi^k = L_\beta(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{\lambda}^k) + \frac{6L^2}{\mu^2\beta}\|\boldsymbol{y}^k - \boldsymbol{y}^{k-1}\|^2. \tag{52}$$

Then after running SPIDER-ADMM by $K = \mathcal{O}(\epsilon^{-2})$ iterations, we find an $\mathcal{O}(\epsilon)$-approximate KKT point in expectation.

Theorem IV.2 (cont'd)

Specifically, letting $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{\lambda}})$ uniformly randomly taken from $\{\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{\lambda}^k\}_{k=1}^K$, defining $D = \Phi^0 - \min_{k\geq 0} \mathbb{E}[\Phi^k]$, and assuming that D is finite, we have

$$\tilde{\mathbb{E}}\left[\|\mathbf{A}\tilde{\boldsymbol{x}} + \mathbf{B}\tilde{\boldsymbol{y}} - \boldsymbol{b}\|^2\right] \leq \frac{1}{\beta}\left(\frac{D}{K} + \frac{m\sigma^2\eta}{2S_1}\right) = \mathcal{O}(\epsilon^2), \tag{53}$$

$$\tilde{\mathbb{E}}\left[\|\nabla g(\tilde{\boldsymbol{y}}) + \mathbf{B}^T\tilde{\boldsymbol{\lambda}}\|^2\right] \leq \frac{4L^2}{\rho}\left(\frac{D}{K} + \frac{m\sigma^2\eta}{2S_1}\right) = \mathcal{O}(\epsilon^2), \tag{54}$$

$$\begin{aligned}\tilde{\mathbb{E}}\left[\|\nabla f_i(\tilde{\boldsymbol{x}}_i) + \mathbf{A}_i^T\tilde{\boldsymbol{\lambda}}\|^2\right] &\leq 4\frac{K+1}{K}C_i\left(\frac{D}{K+1} + \frac{m\sigma^2\eta}{2S_1}\right) \\ &\quad + \frac{4\sigma^2}{S_1} = \mathcal{O}(\epsilon^2),\end{aligned} \tag{55}$$

where $\{C_i\}_i$ are constants $\tilde{\mathbb{E}}$ denotes taking expectation for all the randomness in NC-SADMM and the selection of $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{\lambda}})$.

Outline

I Stochastic ADMM
SADMM

II The Variation Reduction Technique
SVRG-ADMM

III Fusing VR with Momentum
Acc-SADMM

IV Stochastic Nonconvex Optimization
NC-SADMM
SPIDER-ADMM

References

References

1. Lin, Zhouchen, Huan Li, and Cong Fang. *Alternating Direction Method of Multipliers for Machine Learning.* Springer Nature, 2022.
2. Boyd, Stephen, Stephen P. Boyd, and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.