ADMM for Nonconvex Optimization

Hailiang ZHAO @ ZJU-CS
http://hliangzhao.me

October 27, 2022

Outline

Outline

Multi-Block Bregman ADMM

In this slide, we introduce ADMM for nonconvex problem. Consider the following multi-block linearly constrained problem:

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

under the following assumption:

Assumption I.1

Assume that f_i, $i \in [m]$ are proper lwoer semicontinuous functions and g is L-smooth. Both f_i and g can be nonconvex.

Multi-Block Bregman ADMM

To solve (1), we introduce the following *Multi-Block Bregman ADMM*:

$$
\begin{aligned}
\boldsymbol{x}_i^{k+1} &= \operatorname*{argmin}_{\boldsymbol{x}_i} \Big(f_i(\boldsymbol{x}_i) + \langle \boldsymbol{\lambda}^k, \mathbf{A}_i \boldsymbol{x}_i \rangle \\
&\quad + \frac{\beta}{2} \Big\| \sum_{j<i} \mathbf{A}_j \boldsymbol{x}_j^{k+1} + \mathbf{A}_i \boldsymbol{x}_i + \sum_{j>i} \mathbf{A_j} \boldsymbol{x}_j^k + \mathbf{B} \boldsymbol{y}^k - \boldsymbol{b} \Big\|^2 \\
&\quad + D_{\phi_i}(\boldsymbol{x}_i, \boldsymbol{x}_i^k) \Big), \qquad \forall i \in [m] \text{ in seq.,} \qquad (2) \\
\boldsymbol{y}^{k+1} &= \operatorname*{argmin}_{\boldsymbol{y}} \Big(\mathrm{g}(\boldsymbol{y}) + \langle \boldsymbol{\lambda}^k, \mathbf{B} \boldsymbol{y} \rangle \\
&\quad + \frac{\beta}{2} \Big\| \sum_{i=1}^m \mathbf{A}_i \boldsymbol{x}_i^{k+1} + \mathbf{B} \boldsymbol{y} - \boldsymbol{b} \Big\|^2 + D_{\phi_0}(\boldsymbol{y}, \boldsymbol{y}^k) \Big), \qquad (3) \\
\boldsymbol{\lambda}^{k+1} &= \boldsymbol{\lambda}^k + \beta \Big(\sum_{i=1}^m \mathbf{A}_i \boldsymbol{x}_i^{k+1} + \mathbf{B} \boldsymbol{y}^{k+1} - \boldsymbol{b} \Big). \qquad (4)
\end{aligned}
$$

Multi-Block Bregman ADMM

Similar to LADMM-1 and LADMM-2, we need to choose suitable ϕ_i such that each subproblem can be solved easily. For example, let

$$\phi_0(\boldsymbol{y}) := \frac{L + \beta\|\mathbf{B}\|^2}{2}\|\boldsymbol{y}\|^2 - g(\boldsymbol{y}) - \frac{\beta}{2}\|\mathbf{B}\boldsymbol{y}\|^2, \tag{5}$$

then the $\boldsymbol{y}$ update reduces to

$$\boldsymbol{y}^{k+1} = \boldsymbol{y}^k - \frac{1}{L + \beta\|\mathbf{B}\|^2}\nabla_{\boldsymbol{y}^k} L_\beta(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k, \boldsymbol{\lambda}^k), \tag{6}$$

where L_β is the augmented Lagrangian function of (1), and

$$\boldsymbol{x} = [\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_m^T]^T, \mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_m]. \tag{7}$$

Multi-Block Bregman ADMM

Theorem I.1

Assume that Assumption I.1 and the surjectiveness of $\mathbf{B}$ holds (*there exists* $\sigma > 0$ *such that* $\|\mathbf{B}^T\boldsymbol{\lambda}\| \geq \sigma\|\boldsymbol{\lambda}\|$ *for all* $\boldsymbol{\lambda}$*, it means that* $\mathbf{B}$ *needs to be fully rank*), and ϕ_i is ρ-strongly convex and L_i-smooth with

$$\rho > \frac{12(L^2 + 2L_0^2)}{\sigma^2\beta}, i = 0, ..., m. \tag{8}$$

Suppose that the sequence $\{(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{\lambda}^k)\}_k$ is bounded and $\sum_{i=1}^m f_i(\boldsymbol{x}_i) + g(\boldsymbol{y})$ is bounded below with bounded $(\boldsymbol{x}, \boldsymbol{y})$. Then Multi-Block Bregman ADMM needs $\mathcal{O}(\frac{1}{\epsilon^2})$ iterations to find an ϵ-approximate KKT point $(\boldsymbol{x}^{k+1}, \boldsymbol{y}^{k+1}, \boldsymbol{\lambda}^{k+1})$.

Multi-Block Bregman ADMM

Theorem I.1 (cont'd)

Namely,

$$\Big\| \sum_{i=1}^{m} \mathbf{A}_i \boldsymbol{x}_i^{k+1} + \mathbf{B}\boldsymbol{y}^{k+1} - \boldsymbol{b} \Big\| \leq \mathcal{O}(\epsilon), \tag{9}$$

$$\|\nabla g(\boldsymbol{y}^{k+1}) + \mathbf{B}^T \boldsymbol{\lambda}^{k+1}\| \leq \mathcal{O}(\epsilon), \tag{10}$$

$$\mathrm{dist}\Big(-\mathbf{A}_i^T \boldsymbol{\lambda}^{k+1}, \partial f_i(\boldsymbol{x}_i^{k+1}) \Big) \leq \mathcal{O}(\epsilon), \forall i \in [m]. \tag{11}$$

Pay attention to the conditions to meet for its convergence.

Multi-Block Bregman ADMM

In the proof of Theorem I.1, a crucial step is to bound the dual variables by the primal ones:

$$\sigma^2\|\boldsymbol{\lambda}^{k+1}-\boldsymbol{\lambda}^k\|^2 \le 3(L^2+L_0^2)\|\boldsymbol{y}^{k+1}-\boldsymbol{y}^k\|^2+3L_0^2\|\boldsymbol{y}^k-\boldsymbol{y}^{k-1}\|^2,$$

which is established via the surjectiveness assumption. Nevertheless, we can replace it by $\mathrm{Im}(\mathbf{A}_i)\subseteq\mathrm{Im}(\mathbf{B})$. In this case we have

$$\boldsymbol{\lambda}^{k+1}-\boldsymbol{\lambda}^k=\beta\Big(\sum_{i=1}^m \mathbf{A}_i\boldsymbol{x}_i^{k+1}-\mathbf{B}\boldsymbol{y}^{k+1}-\boldsymbol{b}\Big)\in\mathrm{Im}(\mathbf{B}). \qquad (12)$$

Multi-Block Bregman ADMM

Suppose that the SVD of $\mathbf{B}$ is $\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. Then we may write $\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k = \mathbf{U}\boldsymbol{\alpha}$. Further we have

$$\begin{aligned} \|\mathbf{B}^T(\lambda^{k+1} - \boldsymbol{\lambda}^k)\|^2 &= \|\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{U}\boldsymbol{\alpha}\|^2 = \|\boldsymbol{\Sigma}\boldsymbol{\alpha}\|^2 \\ &\geq \lambda_+(\mathbf{B}\mathbf{B}^T)\|\boldsymbol{\alpha}\|^2 = \lambda_+(\mathbf{B}\mathbf{B}^T)\|\lambda^{k+1} - \boldsymbol{\lambda}^k\|^2, \end{aligned} \tag{13}$$

where $\lambda_+(\mathbf{B}\mathbf{B}^T)$ is the smallest strictly positive eigenvalue of $\mathbf{B}\mathbf{B}^T$. In this case, we do not require $\mathbf{B}$ to be full rank any more.

Besides, when the problem has only one block, i.e., $f_i = 0, \forall i \in [m]$, the assumption $\mathrm{Im}(\mathbf{A}_i) \subseteq \mathrm{Im}(\mathbf{B})$ can be removed since

$$\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k = \beta(\mathbf{B}\boldsymbol{y}^{k+1} - \boldsymbol{b}) \in \mathrm{Im}(\mathbf{B}) \tag{14}$$

always holds.

Multi-Block Bregman ADMM

In Theorem I.1, we only assume that g is smooth, which *allows f_i to be nonsmooth* and can be applied to problems such as sparse and low-rank optimization. Besides, the assumption on the linear constraint, either surjectiveness or $\mathrm{Im}(\mathbf{A}_i) \subseteq \mathrm{Im}(\mathbf{B})$, also plays a critical role.

In the following, we introduce the convergence proof with more assumptions on the objectives instead.

Assumption I.2

All f_i's and g are L-smooth.

We also have convergence result under Assumption I.2.

Multi-Block Bregman ADMM

Theorem I.2

Assume that Assumption I.2 and ϕ_i is ρ-strongly convex and L_i-smooth with

$$\rho > \frac{4\max\{c_1 + c_2, c_3 + c_4\}}{\beta\lambda_+}, i = 0, \dots, m. \tag{15}$$

where λ_+ is the smallest strictly positive eigenvalue of $[\mathbf{A}, \mathbf{B}][\mathbf{A}, \mathbf{B}]^T$, $c_1 \sim c_4$ are specific constants. Suppose that the sequence $\{(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{\lambda}^k)\}_k$ is bounded and $\sum_{i=1}^m f_i(\boldsymbol{x}_i) + g(\boldsymbol{y})$ is bounded below with bounded $(\boldsymbol{x}, \boldsymbol{y})$. Let $\boldsymbol{\lambda}^0 = \mathbf{0}$. Then Multi-Block Bregman ADMM needs $\mathcal{O}(\frac{1}{\epsilon^2})$ iterations to find an ϵ-approximate KKT point $(\boldsymbol{x}^{k+1}, \boldsymbol{y}^{k+1}, \boldsymbol{\lambda}^{k+1})$. Namely, (9), (10), and

$$\|\nabla f_i(\boldsymbol{x}_i^{k+1}) + \mathbf{A}^T\boldsymbol{\lambda}^{k+1}\| \leq \mathcal{O}(\epsilon) \tag{16}$$

hold.

Outline

LADMM-EA

Multi-Block Bregman ADMM uses the Bregman distance $D_{\phi_i}(\boldsymbol{x}_i, \boldsymbol{x}_i^k)$, which results in the proximal term $\frac{\beta'}{2}\|\boldsymbol{x} - \boldsymbol{x}^k\|^2$. In the following, we use $\frac{\beta'}{2}\|\boldsymbol{x} - \boldsymbol{z}^k\|^2$ instead, where $\boldsymbol{z}_i^k$ is an exponential averaging of $\boldsymbol{x}_i^0, \ldots, \boldsymbol{x}_i^k$.
Consider the following problem:

$$\min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\boldsymbol{y}} \sum_{i=1}^{m} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m), \quad s.t. \quad \sum_{i=1}^{m} \mathbf{A}_i \boldsymbol{x}_i = \boldsymbol{b}, \tag{17}$$

with a non-separable objective. Denote

$$\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_m^T)^T \text{ and } \mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_m). \tag{18}$$

LADMM-EA

Consider the following proximal augmented Lagrangian function:

$$P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \mathbf{A}\boldsymbol{x} - \boldsymbol{b} \rangle + \frac{\beta}{2}\|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2. \tag{19}$$

Then we have the following iteration steps:

$$\boldsymbol{x}_j^{k+1} = \boldsymbol{x}_j^k - \alpha_1 \nabla_{\boldsymbol{x}_j} P(\boldsymbol{x}_1^{k+1}, \ldots, \boldsymbol{x}_{j-1}^{k+1}, \boldsymbol{x}_j, \boldsymbol{x}_{j+1}^k, \ldots, \boldsymbol{x}_m^k, \boldsymbol{z}^k, \boldsymbol{\lambda}^k), \tag{20}$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \alpha_2(\mathbf{A}\boldsymbol{x}k - \boldsymbol{b}), \tag{21}$$

$$\boldsymbol{z}^{k+1} = \boldsymbol{z}^k + \alpha_3(\boldsymbol{x}^{k+1} - \boldsymbol{z}^k). \tag{22}$$

We call it Linearized ADMM with Exponential Averaging (*LADMM-EA*).

LADMM-EA

Note that (22) gives

$$\boldsymbol{z}^{k+1} = \sum_{t=0}^{k} \alpha_3 (1-\alpha_3)^{k-t} \boldsymbol{x}^{t+1} + (1-\alpha_3)\boldsymbol{z}^0. \tag{23}$$

LADMM-EA has the following convergence result.

Theorem II.1

Assume that f is L-smooth w.r.t. $\boldsymbol{x}$. Choosing α_1, α_2, and α_3 appropriately and letting $\rho > L$. Then LADMM-EA needs $\mathcal{O}(\frac{1}{\epsilon^2})$ iterations to find an ϵ-appropriate KKT point $(\boldsymbol{x}, \boldsymbol{\lambda})$. Namely,

$$\|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\| \leq \mathcal{O}(\epsilon), \tag{24}$$

$$\|\nabla f(\boldsymbol{x}) + \mathcal{A}^T \boldsymbol{\lambda}\| \leq \mathcal{O}(\epsilon). \tag{25}$$

The result is built on several intermediate lemmas.

Outline

ADMM for Multilinearly Constrained Optimization

ADMM can also be used to solve problems with multilinear constraints in the form of $\mathbf{XY} = \mathbf{Z}$, where multilinear means $\mathbf{XY} = \mathbf{Z}$ is linear w.r.t. the individual variables $\mathbf{X}$ and $\mathbf{Y}$, but nonconvex for $\mathbf{X}$ and $\mathbf{Y}$ jointly. Typical problems include non-negative matrix factorization, RPCA, and the training of neural networks, etc.

More details on this topic can be found at Sec. 4.3 of the ADMM book.

Outline

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.