ADMM without a Fixed Penalty Parameter: Faster Convergence with New Adaptive Penalization

Yi Xu†, Mingrui Liu†, Qihang Lin‡, Tianbao Yang†
†Computer Science Department, ‡Management Sciences Department, The University of Iowa, Iowa City, IA, USA

Background

- Alternating direction method of multipliers (ADMM) has received tremendous interest for solving numerous problems in machine learning, statistics and signal processing.
- The performance of ADMM and many of its variants is very **sensitive to the penalty parameter** of a quadratic penalty applied to the equality constraints.
- Although several approaches have been proposed for dynamically changing this parameter, they do not yield theoretical improvement in the convergence rate and are not directly applicable to stochastic ADMM.

Structured Non-smooth and Non-strongly Convex Problem

The optimization problem of interest:

$$\min_{\mathbf{x}\in\Omega} \quad F(\mathbf{x}) \triangleq f(\mathbf{x}) + \psi(A\mathbf{x}) \tag{1}$$

where $\Omega \subseteq \mathbb{R}^d$ is a closed convex set, $f : \mathbb{R}^d \to \mathbb{R}$ and $\psi : \mathbb{R}^m \to \mathbb{R}$ are proper lower-semicontinuous convex functions, and $A \in \mathbb{R}^{m\times d}$ is a matrix. To apply ADMM, the problem (1) can be cast into the following equivalent constrained optimization problems.

- The optimization problem in **deterministic setting**:

$$\min_{\mathbf{x}\in\Omega,\mathbf{y}\in\mathbb{R}^m} \quad F(\mathbf{x}) \triangleq f(\mathbf{x}) + \psi(\mathbf{y}), \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x}. \tag{2}$$

- The optimization problem in **stochastic setting**:

$$\min_{\mathbf{x}\in\Omega,\mathbf{y}\in\mathbb{R}^m} \quad F(\mathbf{x}) \triangleq \mathrm{E}_\xi[f(\mathbf{x};\xi)] + \psi(\mathbf{y}), \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x}. \tag{3}$$

where ξ is a random variable.

Let Ω_*, F_* denote the set of optimal solutions and the optimal value, respectively.

- We make the following assumptions:
  - a. There exist $x_0 \in \Omega$ and $\epsilon_0 \geq 0$ s.t. $F(x_0) - F_* \leq \epsilon_0$;
  - b. Ω_* is a non-empty convex compact set;
  - c. There exists $\rho > 0$ s.t. $\|\partial\psi(\mathbf{y})\|_2 \leq \rho$ for all $\mathbf{y}$;
  - d. ψ is defined everywhere;
  - e. There exists $R > 0$ s.t. $\|\partial f(\mathbf{x};\xi)\|_2 \leq R$ almost surely for any $\mathbf{x} \in \Omega$ (for stochastic setting only) .
- The ϵ-sublevel set of $F(x)$: $\mathcal{S}_\epsilon = \{x \in \Omega_1 : F(x) \leq F_* + \epsilon\}$
- The distance of x to Ω_*: $dist(x, \Omega_*) = \min_{z\in\Omega_*} \|z - x\|_2$
- The closest point on the $\mathcal{S}_\epsilon$ to x: $x_\epsilon^\dagger = \arg\min_{z\in\mathcal{S}_\epsilon} \|z - x\|_2^2$

Alternating Direction Method of Multipliers (ADMM)

An augmented Lagrangian function for (2):

$$L(\mathbf{x}, \mathbf{y}, \lambda) = f(\mathbf{x}) + \psi(\mathbf{y}) - \lambda^\top(A\mathbf{x} - \mathbf{y}) + \frac{\beta}{2}\|A\mathbf{x} - \mathbf{y}\|_2^2, \tag{4}$$

where β: penalty parameter, $\lambda \in \mathbb{R}^m$: dual variable.

(1) The **standard ADMM** solves problem (2) iteratively:

$$\mathbf{x}_{\tau+1} = \arg\min_{\mathbf{x}\in\Omega} L(\mathbf{x}, \mathbf{y}_\tau, \lambda_\tau) = \arg\min_{\mathbf{x}\in\Omega} f(\mathbf{x}) + \frac{\beta}{2}\left\|(A\mathbf{x} - \mathbf{y}_\tau) - \frac{1}{\beta}\lambda_\tau\right\|_2^2, \tag{5}$$

$$\mathbf{y}_{\tau+1} = \arg\min_{\mathbf{x}\in\Omega} L(\mathbf{x}_{\tau+1}, \mathbf{y}, \lambda_\tau) = \arg\min_{\mathbf{y}\in\mathbb{R}^m} \psi(\mathbf{y}) + \frac{\beta}{2}\left\|(A\mathbf{x}_{\tau+1} - \mathbf{y}) - \frac{1}{\beta}\lambda_\tau\right\|_2^2, \tag{6}$$

$$\lambda_{\tau+1} = \lambda_\tau - \beta(A\mathbf{x}_{\tau+1} - \mathbf{y}_{\tau+1}). \tag{7}$$

- When $A \neq I$, solving the subproblem (5) might be difficult. To alleviate the issue, **linearized ADMM** solves the following problem instead of (5):

$$\mathbf{x}_{\tau+1} = \arg\min_{\mathbf{x}\in\Omega} f(\mathbf{x}) + \frac{\beta}{2}\left\|(A\mathbf{x} - \mathbf{y}_\tau) - \frac{1}{\beta}\lambda_\tau\right\|_2^2 + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_\tau\|_G^2, \tag{8}$$

where $\|\mathbf{x}\|_G = \sqrt{\mathbf{x}^\top G\mathbf{x}}$ and $G \in \mathbb{R}^{d\times d}$ is a positive semi-definite matrix.

- By setting $G = \gamma I - \beta A^\top A \succeq 0$, the term $\mathbf{x}^\top A^\top A\mathbf{x}$ in (8) vanishes.

Algorithm 1 ADMM $(\mathbf{x}_0, \beta, t)$

1: **Input**: $\mathbf{x}_0 \in \Omega$, β, t.
2: **Initialize:** $\mathbf{x}_1 = \mathbf{x}_0, \mathbf{y}_1 = A\mathbf{x}_1, \lambda_1 = 0$, $\gamma = \beta\|A\|_2^2$ and $G = \gamma I - \beta A^\top A$ or $G = 0$.
3: **for** $\tau = 1, \ldots, t$ **do**
4: Update $\mathbf{x}_{\tau+1}$ by (8), $\mathbf{y}_{\tau+1}$ by (6), $\lambda_{\tau+1}$ by (7)
5: **end for**
6: **Output**: $\widehat{\mathbf{x}}_t = \sum_{\tau=1}^t \mathbf{x}_\tau/t$

Lemma 1 [1]. *By setting $\beta = \frac{\rho}{\sqrt{2}\|A\|_2\|\mathbf{x}_*-\mathbf{x}_0\|_2}$ (β depends on unknown $\mathbf{x}_*$), after $t = O(1/\epsilon)$ iterations, ADMM ensures that*

$$F(\widehat{\mathbf{x}}_t) - F(\mathbf{x}_*) \leq \epsilon.$$

(2) The **stochastic ADMM** updates $\mathbf{y}_{\tau+1}$ and $\lambda_{\tau+1}$ the same to (6) and (7), but updates $\mathbf{x}_{\tau+1}$ as

$$\mathbf{x}_{\tau+1} = \arg\min_{\mathbf{x}\in\Omega} f(\mathbf{x}_\tau;\xi_\tau) + \partial f(\mathbf{x}_\tau;\xi_\tau)^\top(\mathbf{x}-\mathbf{x}_\tau) + \frac{\beta}{2}\left\|(A\mathbf{x} - \mathbf{y}_\tau) - \frac{1}{\beta}\lambda_\tau\right\|_2^2 + \frac{\|\mathbf{x} - \mathbf{x}_\tau\|_{G_\tau}^2}{\eta_\tau} \tag{9}$$

where η_τ is a stepsize and $G_\tau = \gamma I - \beta\eta_\tau A^\top A \succeq I$ or $G_\tau = I$.

Algorithm 2 SADMM $(\mathbf{x}_0, \eta, \beta, t, \Omega)$

1: **Input**: $\mathbf{x}_0 \in \mathbb{R}^d$, η, β, t.
2: **Initialize:** $\mathbf{x}_1 = \mathbf{x}_0, \mathbf{y}_1 = A\mathbf{x}_1, \lambda_1 = 0$, $G_\tau = \gamma I - \eta\beta A^\top A \succeq I$
3: **for** $\tau = 1, \ldots, t$ **do**
4: Update $\mathbf{x}_{\tau+1}$ by (9), $\mathbf{y}_{\tau+1}$ by (6), $\lambda_{\tau+1}$ by (7)
5: **end for**
6: **Output**: $\widehat{\mathbf{x}}_t = \sum_{\tau=1}^t \mathbf{x}_\tau/t$

Lemma 2 [2]. *By setting $\beta = \frac{\rho}{\|A\|_2\|\mathbf{x}_*-\mathbf{x}_0\|_2}$ (β depends on unknown $\mathbf{x}_*$), after $t = O(1/\epsilon^2)$ iterations, with high probability, SADMM ensures that*

$$F(\widehat{\mathbf{x}}_t) - F(\mathbf{x}) \leq \epsilon.$$

Local Error Bound and Global Error Inequality

Definition 1. *A function $F(\mathbf{x})$ is said to satisfy a local error bound condition on ϵ-sublevel set if there exist $\theta \in (0, 1]$ and $c > 0$ such that for any $\mathbf{x} \in \mathcal{S}_\epsilon$*

$$dist(\mathbf{x}, \Omega_*) \leq c(F(\mathbf{x}) - F_*)^\theta. \tag{10}$$

Lemma 3 [3]. *For any $\mathbf{x} \in \Omega$ and $\epsilon > 0$, we have*

$$\|\mathbf{x} - \mathbf{x}_\epsilon^\dagger\|_2 \leq \frac{dist(\mathbf{x}_\epsilon^\dagger, \Omega_*)}{\epsilon}(F(\mathbf{x}) - F(\mathbf{x}_\epsilon^\dagger)) \tag{11}$$

where $\mathbf{x}_\epsilon^\dagger \in \mathcal{S}_\epsilon$ is the closest point in the ϵ-sublevel set to $\mathbf{x}$.

Locally Adaptive ADMM (LA-ADMM)

Algorithm 3 LA-ADMM $(\mathbf{x}_0, \beta_1, K, t)$

1: **Input**: $\mathbf{x}_0 \in \Omega$, K, t, initial β_1
2: **for** $k = 1, \ldots, K$ **do**
3: Let $\mathbf{x}_k = \text{ADMM}(\mathbf{x}_{k-1}, \beta_k, t)$
4: Update $\beta_{k+1} = 2\beta_k$
5: **end for**
6: **Output**: $\mathbf{x}_K$

Main Result 1

Theorem 1. Assume $F(x)$ obeys the local error bound condition. Let LA-ADMM run with $t = O\left(\left\lceil\frac{8\rho\|A\|_2\max(1,c^2)}{\epsilon^{1-\theta}}\right\rceil\right)$ iterations for each stage and $K = \lceil\log_2(\epsilon_0/\epsilon)\rceil$ with $\beta_1 = \frac{2\rho\epsilon^{1-\theta}}{\|A\|_2\epsilon_0}$. Then $F(x_K) - F_* \leq 2\epsilon$. Hence, the iteration complexity of LA-ADMM is $\widetilde{O}(1/\epsilon^{1-\theta})$.

- Remark 1. The number of iteration t depends on the unknown parameter c. This dependence can be relaxed by using another level of restarting and increasing sequence of t. We refer readers to our paper for more details.

Locally Adaptive Stochastic ADMM (LA-SADMM)

Algorithm 4 LA-SADMM $(\mathbf{x}_0, \eta_1, \beta_1, D_1, K, t)$

1: **Input**: $\mathbf{x}_0 \in \mathbb{R}^d$, K, t, η_1, initial β_1 and D_1.
2: **for** $k = 1, \ldots, K$ **do**
3: Let $\mathbf{x}_k = \text{SADMM}(\mathbf{x}_{k-1}, \eta_k, \beta_k, t, \mathcal{B}_k \cap \Omega)$
4: Update $\eta_{k+1} = \eta_k/2$ and $\beta_{k+1} = 2\beta_k$, $D_{k+1} = D_k/2$.
5: **end for**
6: **Output**: $\mathbf{x}_K$

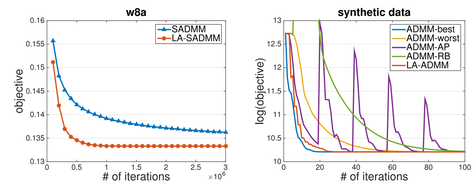
Main Result 2

Theorem 2. Assume $F(x)$ obeys the local error bound condition. Given $\delta \in (0, 1)$ and $\tilde{\delta} = \delta/K$, let LA-SADMM run with $t \geq \max\{\frac{6912R^2\log(1/\tilde{\delta})D_1^2}{\epsilon_0^2}, \frac{12\rho\|A\|_2D_1}{\epsilon_0}, \frac{\rho^2\|A\|_2^2}{R^2}\}$ iterations for each stage and $K = \lceil\log_2(\frac{\epsilon_0}{\epsilon})\rceil$ with $\eta_1 = \frac{\epsilon_0}{6R^2}$, $\beta_1 = \frac{6R^2}{\|A\|_2^2\epsilon_0}$, $D_1 \geq \frac{c\epsilon_0}{\epsilon^{1-\theta}}$ and $G_\tau = 2I - \eta_1\beta_1 A^\top A \succeq I$. Then $F(\mathbf{x}_K) - F_* \leq 2\epsilon$ with probability $1 - \delta$. Hence, the iteration complexity of LA-SADMM with probability $1 - \delta$ is $\widetilde{O}(\log(1/\delta)/\epsilon^{2(1-\theta)})$.

- Remark 2. The radius D_1 depends on the unknown parameter c. This dependence can be relaxed by using another level of restarting and increasing sequence of t. We refer readers to our paper for more details.

Applications and Experiments

1. Generalized LASSO: $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{x}^\top\mathbf{a}_i, b_i) + \delta\|A\mathbf{x}\|_1$
   - $(\mathbf{a}_i, b_i)$: a set of pairs of training data, $i = 1, \ldots, n$; $\delta \geq 0$: regularization parameter; $A \in \mathbb{R}^{m\times d}$: specified matrix; $\ell(z, b)$: convex loss function in terms of z.
   - Different types of LASSO:
     - Standard LASSO: $A = I \in \mathbb{R}^{d\times d}$
     - Fused LASSO: penalizes ℓ_1 norm of coefficients and successive differences
     - Graph-guided fused LASSO (GGLASSO): $A \in \mathbb{R}^{m\times d}$ encodes graph information
     - Sparse graph-guided fused LASSO (S-GGLASSO): $\|A\mathbf{x}\|_1 = \delta_2\|\mathbf{x}\|_1 + \delta_1\|F\mathbf{x}\|_1$
   - Piecewise linear loss:
     - hinge loss $\ell(z, b) = \max(0, 1 - bz)$, absolute loss $\ell(z, b) = |z - b|$, ϵ-insensitive loss $\ell(z, b) = \max(|z - b| - \epsilon, 0)$
     - $\theta = 1$: both LA-ADMM and LA-SADMM achieve linear convergence $O(\log(1/\epsilon))$
   - Piecewise quadratic loss:
     - square loss $\ell(z, b) = (z - b)^2$; squared hinge loss $\ell(z, b) = \max(0, 1 - bz)^2$
     - $\theta = 1/2$: LA-ADMM and LA-SADMM achieve iteration complexities of $\widetilde{O}(1/\sqrt{\epsilon})$ and $\widetilde{O}(1/\epsilon)$, respectively
2. Robust Regression with a Low-rank Regularizer: $F(X) = \lambda\|X\|_* + \|AX - C\|_1$
3. Low-rank Representation: $F(X) = \lambda\|X\|_* + \|AX - A\|_{2,1}$.



(a) SVM + GGLASSO (b) RR + LR

(c) SVM + S-GGLASSO (d) LRR

[1] Bingsheng He and Xiaoming Yuan. On the $O(1/n)$ convergence rate of the douglas-rachford alternating direction method. SIAM J. Numer. Anal., 50(2):700-709, 2012.
[2] Taiji Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. ICML, pages 392-400, 2013.
[3] Tianbao Yang and Qihang Lin. RSG: Beating subgradient method without smoothness and strong convexity. CoRR, abs/1512.03107, 2016.