Supplementary Material: Adaptive Consensus ADMM for Distributed Optimization

Zheng Xu [1] **Gavin Taylor** [2] **Hao Li** [1] **Mário A. T. Figueiredo** [3] **Xiaoming Yuan** [4] **Tom Goldstein** [1]

This is the supplemental material for Adaptive Consensus ADMM (ACADMM) (Xu et al., 2017c). We provide details of proofs and experimental settings, in addition to more results. Our proof generalizes the variational inequality approach in (He et al., 2000; He & Yuan, 2012; 2015; Xu et al., 2017b).

1. Proof of lemmas

1.1. Proof of Lemma 1 (17)

Proof. By using the updated dual variable λ^{k+1} in (10), VI (15) can be rewritten as

$$\forall v,\, g(v) - g(v^{k+1}) - (Bv - Bv^{k+1})^T \lambda^{k+1} \geq 0. \tag{S1}$$

Similarly, in the previous iteration,

$$\forall v,\, g(v) - g(v^{k}) - (Bv - Bv^{k})^T \lambda^{k} \geq 0. \tag{S2}$$

Let $v = v^k$ in (S1) and $v = v^{k+1}$ in (S2), and sum the two inequalities together. We conclude

$$(Bv^{k+1} - Bv^k)^T(\lambda^{k+1} - \lambda^k) \geq 0. \tag{S3}$$

□

1.2. Proof of Lemma 1 (18)

Proof. VI (16) can be rewritten as

$$\phi(y) - \phi(y^{k+1}) + (z - z^{k+1})^T \left(F(z^{k+1}) + \Omega(\Delta z_k^+, T^k)\right) \geq 0, \tag{S4}$$

where $\Omega(\Delta z_k^+, T^k) = (-A^T T^k B \Delta v_k^+; 0; (T^k)^{-1}\Delta\lambda_k^+)$.

Let $y = y^*, z = z^*$ in VI (S4), and $y = y^{k+1}, z = z^{k+1}$ in VI (13), and sum the two equalities together to get

$$(\Delta z_{k+1}^*)^T \Omega(\Delta z_k^+, T^k) \geq (\Delta z_{k+1}^*)^T (F(z^*) - F(z^{k+1})). \tag{S5}$$

Since $F(z)$ is monotone, the right hand side is non-negative. Now, substitute $\Omega(\Delta z_k^+, T^k)$ into (S5) to get

$$-(A\Delta u_{k+1}^*)^T T^k (B\Delta v_k^+) + (\Delta\lambda_{k+1}^*)^T (T^k)^{-1} \Delta\lambda_k^+ \geq 0. \tag{S6}$$

If we use the feasibility constraint of optimal solution ($Au^* + Bv^* = b$) and the dual update formula (10), we have

$$T^k A \Delta u_{k+1}^* = \Delta\lambda_k^+ - T^k B \Delta v_{k+1}^*. \tag{S7}$$

Substitute this into (S6) yields

$$(B\Delta v_{k+1}^*)^T T^k B \Delta v_k^+ + (\Delta\lambda_{k+1}^*)^T (T^k)^{-1} \Delta\lambda_k^+ \geq (B\Delta v_k^+)^T \Delta\lambda_k^+ \tag{S8}$$

The proof (18) is concluded by applying (17) to (S8). □

1.3. Proof of Lemma 1 (19)

Proof.

$$\begin{aligned}
\|\Delta z_k^*\|_{H^k}^2 &= \|z^* - z^k\|_{H^k}^2 && \text{(S9)}\\
&= \|z^* - z^{k+1} + z^{k+1} - z^k\|_{H^k}^2 && \text{(S10)}\\
&= \|\Delta z_{k+1}^* + \Delta z_k^+\|_{H^k}^2 && \text{(S11)}\\
&= \|\Delta z_{k+1}^*\|_{H^k}^2 + \|\Delta z_k^+\|_{H^k}^2 + 2(\Delta z_{k+1}^*)^T H^k \Delta z_k^+ && \text{(S12)}\\
&\geq \|\Delta z_{k+1}^*\|_{H^k}^2 + \|\Delta z_k^+\|_{H^k}^2. && \text{(S13)}
\end{aligned}$$

Eq. (18) is used for the inequality in (S13), and Eq. (19) is derived by rearranging the order of $\|\Delta z_k^*\|_{H^k}^2 \geq \|\Delta z_{k+1}^*\|_{H^k}^2 + \|\Delta z_k^+\|_{H^k}^2$. □

1.4. Proof of Lemma 2

Proof. Applying the observation

$$(a-b)^T H (c-d) = \frac{1}{2}(\|a-d\|_H^2 - \|a-c\|_H^2) + \frac{1}{2}(\|c-b\|_H^2 - \|c-d\|_H^2), \tag{S14}$$

[1]University of Maryland, College Park [2]United States Naval Academy, Annapolis, [3]Universidade de Lisboa, Portugal [4]Hong Kong Baptist University, Hong Kong. Correspondence to: Zheng Xu <xuzhustc@gmail.com>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

we have

$$(\tilde{z}^{k+1}-z)^T H^k \Delta z_k^+ = (\tilde{z}^{k+1}-z)H^k(z^{k+1}-z^k) \quad \text{(S15)}$$

$$=\frac{1}{2}(\|\tilde{z}^{k+1}-z^k\|_{H^k}^2 - \|\tilde{z}^{k+1}-z^{k+1}\|_{H^k}^2)+ \frac{1}{2}(\|z^{k+1}-z\|_{H^k}^2 - \|z^k - z\|_{H^k}^2). \quad \text{(S16)}$$

We now consider

$$\|\tilde{z}^{k+1}-z^{k+1}\|_{H^k}^2 = \|\tilde{z}^{k+1}-z^k+z^k-z^{k+1}\|_{H^k}^2 \quad \text{(S17)}$$

$$=\|\tilde{z}^{k+1}-z^k\|_{H^k}^2 + \|\Delta z_k^+\|_{H^k}^2 - 2(\tilde{z}^{k+1}-z^k)^T H^k \Delta z_k^+, \quad \text{(S18)}$$

and get

$$\|\tilde{z}^{k+1}-z^k\|_{H^k}^2 - \|\tilde{z}^{k+1}-z^{k+1}\|_{H^k}^2 \quad \text{(S19)}$$

$$=2(\tilde{z}^{k+1}-z^k)^T H^k \Delta z_k^+ - \|\Delta z_k^+\|_{H^k}^2. \quad \text{(S20)}$$

We then substitute Δz_k^+ with $M^k(\tilde{z}^{k+1}-z^k)$ in (12),

$$\|\tilde{z}^{k+1}-z^k\|_{H^k}^2 - \|\tilde{z}^{k+1}-z^{k+1}\|_{H^k}^2 \quad \text{(S21)}$$

$$=(\tilde{z}^{k+1}-z^k)^T(2I-M^k)^T H^k M^k(\tilde{z}^{k+1}-z^k) \quad \text{(S22)}$$

$$=\|\hat{\lambda}^{k+1}-\lambda^k\|_{(T^k)^{-1}}^2 \geq 0. \quad \text{(S23)}$$

Combining (S16) and (S23), we conclude

$$(\tilde{z}^{k+1}-z)^T H^k \Delta z_k^+ \geq \frac{1}{2}(\|z^{k+1}-z\|_{H^k}^2 - \|z^k-z\|_{H^k}^2). \quad \text{(S24)}$$

□

1.5. Proof of Lemma 3

Proof. Assumption 1 implies (22), which suggests the diagonal matrices T^k and T^{k-1} satisfy

$$T^k \leq (1+(\eta^k)^2)T^{k-1}$$
$$(T^k)^{-1} \leq (1+(\eta^k)^2)(T^{k-1})^{-1}. \quad \text{(S25)}$$

Then we have

$$\|z-z'\|_{H^k}^2 \quad \text{(S26)}$$

$$=\|B(v-v')\|_{T^k}^2 + \|\lambda-\lambda'\|_{(T^k)^{-1}}^2 \quad \text{(S27)}$$

$$\leq(1+(\eta^k)^2)(\|B(v-v')\|_{T^{k-1}}^2 + \|\lambda-\lambda'\|_{(T^{k-1})^{-1}}^2) \quad \text{(S28)}$$

$$\leq(1+(\eta^k)^2)\|z-z'\|_{H^{k-1}}^2. \quad \text{(S29)}$$

The inequality (S25) is used to get from (S27) to (S28). □

1.6. Proof of Lemma 4

Proof. From (27) we know

$$\|\Delta z_k^+\|_{H^k}^2 + \|\Delta z_{k+1}^*\|_{H^k}^2 \leq (1+(\eta^k)^2)\|\Delta z_k^*\|_{H^{k-1}}^2. \quad \text{(S30)}$$

Hence

$$\|\Delta z_{k+1}^*\|_{H^k}^2 \leq (1+(\eta^k)^2)\|\Delta z_k^*\|_{H^{k-1}}^2 \quad \text{(S31)}$$

$$\leq \prod_{t=1}^{k}(1+(\eta^t)^2)\|\Delta z_1^*\|_{H^0}^2 \quad \text{(S32)}$$

$$\leq \prod_{t=1}^{\infty}(1+(\eta^t)^2)\|\Delta z_1^*\|_{H^0}^2 \quad \text{(S33)}$$

$$= C_\eta^\Pi \|\Delta z_1^*\|_{H^0}^2 < \infty. \quad \text{(S34)}$$

Let $z' = z^*$ in Lemma 3, we have

$$\|z-z^*\|_{H^k}^2 \leq (1+(\eta^k)^2)\|z-z^*\|_{H^{k-1}}^2 \quad \text{(S35)}$$

$$\leq \prod_{t=1}^{k}(1+(\eta^t)^2)\|z-z^*\|_{H^0}^2 \quad \text{(S36)}$$

$$\leq \prod_{t=1}^{\infty}(1+(\eta^t)^2)\|z-z^*\|_{H^0}^2 \quad \text{(S37)}$$

$$= C_\eta^\Pi \|z-z^*\|_{H^0}^2 < \infty. \quad \text{(S38)}$$

Let $z' = z^k$ in Lemma 3, we have

$$\|z-z^k\|_{H^k}^2 \leq (1+(\eta^k)^2)\|z-z^k\|_{H^{k-1}}^2. \quad \text{(S39)}$$

Then we have

$$\sum_{k=1}^{l}(\|z-z^k\|_{H^k}^2 - \|z-z^k\|_{H^{k-1}}^2) \quad \text{(S40)}$$

$$\leq \sum_{k=1}^{l}(\eta^k)^2\|z-z^k\|_{H^{k-1}}^2 \quad \text{(S41)}$$

$$= \sum_{k=1}^{l}(\eta^k)^2\|z-z^*+z^*-z^k\|_{H^{k-1}}^2 \quad \text{(S42)}$$

$$\leq \sum_{k=1}^{l}(\eta^k)^2(\|z-z^*\|_{H^{k-1}}^2 + \|\Delta z_k^*\|_{H^{k-1}}^2) \quad \text{(S43)}$$

$$\leq \sum_{k=1}^{l}(\eta^k)^2(C_\eta^\Pi \|z-z^*\|_{H^0}^2 + C_\eta^\Pi \|\Delta z_1^*\|_{H^0}^2) \quad \text{(S44)}$$

$$\leq \sum_{k=1}^{\infty}(\eta^k)^2(C_\eta^\Pi \|z-z^*\|_{H^0}^2 + C_\eta^\Pi \|\Delta z_1^*\|_{H^0}^2) \quad \text{(S45)}$$

$$=C_\eta^\Sigma(C_\eta^\Pi \|z-z^*\|_{H^0}^2 + C_\eta^\Pi \|\Delta z_1^*\|_{H^0}^2) \quad \text{(S46)}$$

$$=C_\eta^\Sigma C_\eta^\Pi(\|z-z^*\|_{H^0}^2 + \|\Delta z_1^*\|_{H^0}^2) < \infty. \quad \text{(S47)}$$

□

1.7. Proof of equivalence of generalized ADMM and DRS in Section 5.1

Proof. The optimality condition for ADMM step (8) is

$$0 \in \partial f(u^{k+1}) - A^T(\underbrace{\lambda^k + T^k(b - Au^{k+1} - Bv^k)}_{\hat{\lambda}^{k+1}}), \quad \text{(S48)}$$

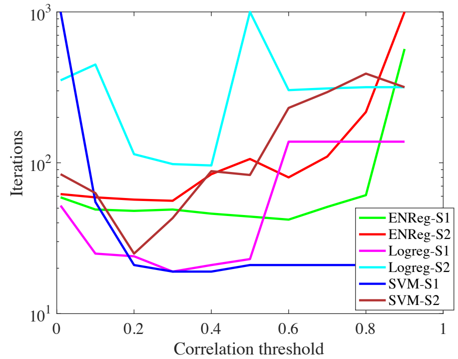

Figure 1: ACADMM is robust to the correlation threshold hyper-parameter ϵ^{cor}.

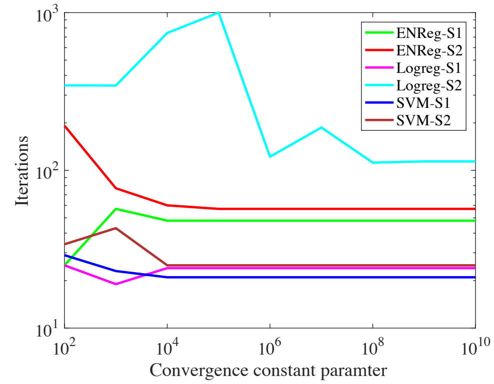

Figure 2: ACADMM is robust to the convergence threshold C_{cg}.

which is equivalent to $A^T\hat{\lambda}^{k+1} \in \partial f(u^{k+1})$. By exploiting properties of the Fenchel conjugate (Rockafellar, 1970), we get $u^{k+1} \in \partial f^*(A^T\hat{\lambda}^{k+1})$. A similar argument using the optimality condition for (9) leads to $v^{k+1} \in \partial g^*(B^T\lambda^{k+1})$. Recalling the definition of $\hat{f}, \hat{g}$ in (42), we arrive at

$$Au^{k+1} - b \in \partial\hat{f}(\hat{\lambda}^{k+1}) \text{ and } Bv^{k+1} \in \partial\hat{g}(\lambda^{k+1}). \quad \text{(S49)}$$

We can then use simple algebra to verify $\lambda^k, \hat{\lambda}^{k+1}$ in (10) and $\partial\hat{f}(\hat{\lambda}^{k+1}), \partial\hat{g}(\lambda^{k+1})$ in (S49) satisfy the generalized DRS steps (43, 44). □

1.8. Proposition for proof in Section 5.2

Proposition 1 (Spectral DRS (Xu et al., 2017a)). *Suppose the Douglas-Rachford splitting steps are used,*

$$0 \in (\hat{\lambda}^{k+1} - \lambda^k)/\tau^k + \partial\hat{f}(\hat{\lambda}^{k+1}) + \partial\hat{g}(\lambda^k) \quad \text{(S50)}$$

$$0 \in (\lambda^{k+1} - \lambda^k)/\tau^k + \partial\hat{f}(\hat{\lambda}^{k+1}) + \partial\hat{g}(\lambda^{k+1}), \quad \text{(S51)}$$

and assume the subgradients are locally linear,

$$\partial\hat{f}(\hat{\lambda}) = \alpha\,\hat{\lambda} + \Psi \quad \textit{and} \quad \partial\hat{g}(\lambda) = \beta\,\lambda + \Phi, \quad \text{(S52)}$$

where $\alpha, \beta \in \mathbb{R}$, $\Psi, \Phi \subset \mathbb{R}^p$. Then, the minimal residual of $\hat{f}(\lambda^{k+1}) + \hat{g}(\lambda^{k+1})$ is obtained by setting $\tau^k = 1/\sqrt{\alpha\,\beta}$.

2. More experimental results

We provide more experimental results demonstrating the robustness of ACADMM in Fig. 1, Fig. 2 and Fig. 3.

3. Synthetic problems in experiments

We provide the details of the synthetic data used in our experiments.

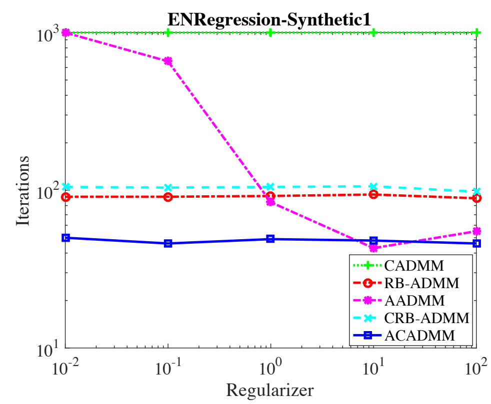

Figure 3: ACADMM is robust to regularizer parameter ρ in EN regression problem.

3.1. Sampling data matrices from Gaussian(s)

For *Synthetic1*, on each compute node i, we create a data matrix $D_i \in \mathbb{R}^{n_i \times d}$ with n_i samples and d features using a standard normal distribution. For *Synthetic2*, we build 10 Gaussian feature sets $\{D_i\}$. On each node, we then randomly choose an index j_i, and randomly select two Gaussian parameters $\mu_1, \ldots, \mu_{10} \in \mathbb{R}$ and $\sigma_1, \ldots, \sigma_{10} \in \mathbb{R}$. We then introduce heterogeneity across nodes by computing

$$D_i \leftarrow D_i * \sigma_{j_i} + \mu_{j_i}. \quad \text{(S53)}$$

3.2. Correlation for Elastic Net regression

Following standard method used to test elastic net regression in (Zou & Hastie, 2005), we introduce correlations into the datasets. We start by building a random Gaussian dataset D_i on each node. We then select the number of active features as $0.6d$. Then we randomly select three

Gaussian vectors $v_{i,1}, v_{i,2}, v_{i,3} \in \mathbb{R}^{n_i}$. We then compute

$$\begin{aligned}\forall j \in & \{1, 2, \dots, 0.2d\}, \\ & D_i[:, j] \leftarrow D_i[:, j] + v_{i,1},\end{aligned} \tag{S54}$$

$$\begin{aligned}\forall j \in & \{0.2d + 1, 0.2d + 2, \dots, 0.4d\}, \\ & D_i[:, j] \leftarrow D_i[:, j] + v_{i,2},\end{aligned} \tag{S55}$$

$$\begin{aligned}\forall j \in & \{0.4d + 1, 0.4d + 2, \dots, 0.6d\}, \\ & D_i[:, j] \leftarrow D_i[:, j] + v_{i,3},\end{aligned} \tag{S56}$$

where $D_i[:, j]$ denotes the jth column of D_i.

3.3. Regression measurement

We use a groundtruth vector $x \in \mathbb{R}^d$, where the first $0.6d$ features are 1 and the rest are 0, and generate measurements for the regression problem as

$$D_i x = c_i \tag{S57}$$

where D_i is random Gaussian.

3.4. Classification labels

For classification problems, we add a constant d_{const} to the active features on half of the feature vectors stored on each node. This means we compute

$$D_i[0.5n_i : n_i,\ 1 : 0.6d] \leftarrow D_i[0.5n_i : n_i,\ 1 : 0.6d] + d_{\text{const}}.$$

We then create a ground truth label vector $c_i \in \mathbb{R}^{n_i}$, which contains 1 for the permuted feature vectors, and -1 for the rest.

References

He, Bingsheng and Yuan, Xiaoming. On the o(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2): 700–709, 2012.

He, Bingsheng and Yuan, Xiaoming. On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. *Numerische Mathematik*, 130: 567–577, 2015.

He, Bingsheng, Yang, Hai, and Wang, Shengli. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Jour. Optim. Theory and Appl.*, 106(2):337–356, 2000.

Rockafellar, R. *Convex Analysis*. Princeton University Press, 1970.

Xu, Zheng, Figueiredo, Mario AT, and Goldstein, Thomas. Adaptive ADMM with spectral penalty parameter selection. *AISTATS*, 2017a.

Xu, Zheng, Figueiredo, Mario AT, Yuan, Xiaoming, Studer, Christoph, and Goldstein, Thomas. Adaptive relaxed ADMM: Convergence theory and practical implementation. *CVPR*, 2017b.

Xu, Zheng, Taylor, Gavin, Li, Hao, Figueiredo, Mario AT, Yuan, Xiaoming, and Goldstein, Tom. Adaptive consensus ADMM for distributed optimization. *ICML*, 2017c.

Zou, Hui and Hastie, Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005.