Non-smooth M-estimator for Maximum Consensus Estimation Supplementary Material

Huu Le, Anders Eriksson, Michael Milford
{first-name}.{last-name}@qut.edu.au
School of Electrical Engineering and Computer Science, Queensland University of Technology, Australia.

Thanh-Toan Do, Tat-Jun Chin
{first-name}.{last-name}@adelaide.edu.au
School of Computer Science, The University of Adelaide, Australia.

David Suter
d.suter@ecu.edu.au
School of Science, Edith Cowan University, Australia.

1 Detailed procedure for updating $\boldsymbol{\theta}_i$

1.1 Quadratic program with one constraint (QCQP1)

Consider the following quadratic optimization problem:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^d} \quad & \|\mathbf{x}-\mathbf{z}\|_2^2, \\ \text{subject to} \quad & \mathbf{x}^T\mathbf{Q}\mathbf{x}+\mathbf{q}^T\mathbf{x}+r\leq 0. \end{aligned} \tag{1}$$

The fact that (1) contains only one quadratic constraint makes it tractable to obtain the global solution for (1), regardless of the convexity of the objective function and the constraint. More details about QCQP1 can be found in [1, 4]

Henceforth, we will show that the updating steps for $\boldsymbol{\theta}_i$, i.e., solving the problem (12) in the main paper, boils down to solving QCQP1 problems in the form of (1).

1.2 Solving for $\boldsymbol{\theta}_i$

Recall from the main paper that updating $\boldsymbol{\theta}_i$ can be done by solving the problem

$$\min_{\boldsymbol{\theta}_i} \Phi(f_i(\boldsymbol{\theta}_i)) - \mu\|\boldsymbol{\theta}_i\|^2 + \rho\|\boldsymbol{\theta}_i - \boldsymbol{\theta} + \boldsymbol{\lambda}_i\|^2, \tag{2}$$

which leads us to the three subproblems (13) and (14) in the main paper. In the following, we detail the steps to solve the problem (13) by converting it into the QCQP1 form of (1). The same approach can be applied to the problems (14). To make it easy for the reader to follow, here we rewrite the problem (13) in the main paper:

$$\begin{aligned} \min_{\boldsymbol{\theta}_i} \quad & -\mu\|\boldsymbol{\theta}_i\|^2 + \rho\|\boldsymbol{\theta}_i - \boldsymbol{\theta} + \boldsymbol{\lambda}_i\|^2 \\ \text{subject to} \quad & 0\leq f_i(\boldsymbol{\theta}_i)\leq \varepsilon, \end{aligned} \tag{3}$$

Consider first the objective function of the problem (3)

$$-\mu\|\boldsymbol{\theta}_i\|^2+\rho\|\boldsymbol{\theta}_i-\boldsymbol{\theta}+\boldsymbol{\lambda}_i\|^2, \tag{4}$$

which can be rewritten as

$$(\rho-\mu)\|\boldsymbol{\theta}_i\|^2-2\rho(\boldsymbol{\theta}-\boldsymbol{\lambda}_i)^T\boldsymbol{\theta}_i+\rho\|\boldsymbol{\theta}-\boldsymbol{\lambda}_i\|^2. \tag{5}$$

As we are minizing over $\boldsymbol{\theta}_i$, the term $\rho\|\boldsymbol{\theta}-\boldsymbol{\lambda}_i\|^2$ in (5) can safely be ignored. Consequently, the cost function becomes

$$(\rho-\mu)\|\boldsymbol{\theta}_i\|^2-2\rho(\boldsymbol{\theta}-\boldsymbol{\lambda}_i)^T\boldsymbol{\theta}_i. \tag{6}$$

Here, by the design of the algorithm, $\rho \gg \mu$. Therefore, minimizing (6) is equivalent to minimizing

$$\|\boldsymbol{\theta}_i\|^2-2\frac{\rho}{\rho-\mu}(\boldsymbol{\theta}-\boldsymbol{\lambda}_i)^T\boldsymbol{\theta}_i, \tag{7}$$

which can be further manipulated to put it in the form of (1). Specifically, the objective function for the optimization problem (3) becomes

$$\left\|\boldsymbol{\theta}_i-\frac{\rho}{\rho-\mu}(\boldsymbol{\theta}-\boldsymbol{\lambda}_i)\right\|_2^2. \tag{8}$$

Next, we show how the constraint of the problem (3) can be put in the form of the quadratic constraint in (1). The quadratic constraint in (3) can be written as

$$0\leq\frac{\|\mathbf{a}_i\boldsymbol{\theta}+\mathbf{b}_i\|_2}{\mathbf{c}_i^T\boldsymbol{\theta}+d_i}\leq\varepsilon. \tag{9}$$

As $\|\mathbf{a}_i\boldsymbol{\theta}+\mathbf{b}_i\|_2\geq 0$, (9) can be equivalently written as

$$\|\mathbf{a}_i\boldsymbol{\theta}+\mathbf{b}_i\|_2^2\leq\varepsilon^2(\mathbf{c}_i^T\boldsymbol{\theta}+d_i)^2, \tag{10}$$

which implies the condition that $\mathbf{c}_i^T\boldsymbol{\theta}+d_i>0$.

It can easily be seen that (10) have the form of quadratic constraint (1), where

$$\mathbf{Q}=\mathbf{a}_{i,1}^T\mathbf{a}_{i,1}+\mathbf{a}_{i,2}^T\mathbf{a}_{i,2}-\varepsilon^2\mathbf{c}_i\mathbf{c}_i^T,$$

$$\mathbf{q}=2(b_{i,1}\mathbf{a}_{i,1}^T+b_{i,2}\mathbf{a}_{i,2}^T-\varepsilon^2 d_i\mathbf{c}_i),$$

and

$$r=b_{i,1}^2+b_{i,2}^2-\varepsilon^2 d_i^2,$$

where $\mathbf{a}_{i,1}$ and $\mathbf{a}_{i,2}$ represents the first and second row of $\mathbf{a}_i$, respectively and, similarly, $b_{i,1}$ and $b_{i,2}$ are the first and second element of $\mathbf{b}_i$.

2 Convergence proof for the ADMM iterations

Our problem is a special case of a non-convex and non-smooth problem discussed in [5]. For completeness, this section provides details for the convergence proof, which was outlined in

Section 3.3 of the main paper. First, we introduce some new notations that will be used throughout the proof.

To prevent clutter, collect all the auxiliary variables $\boldsymbol{\theta}_i$ into a vector $\mathbf{x}$, and all the $\boldsymbol{\lambda}_i$ into $\boldsymbol{\lambda}$. Specifically,

$$\mathbf{x} = [\boldsymbol{\theta}_1^T \ \boldsymbol{\theta}_2^T \dots \boldsymbol{\theta}_N^T]^T, \tag{11}$$

$$\boldsymbol{\lambda} = [\boldsymbol{\lambda}_1^T \ \boldsymbol{\lambda}_2^T \dots \boldsymbol{\lambda}_N^T]^T, \tag{12}$$

Let $\mathbf{B}$ be the negative identity matrix of size $Nd \times Nd$, where N is the number of measurements and d is the dimensionality of $\boldsymbol{\theta}$.

$$\mathbf{B} = -\mathbf{I}_{Nd \times Nd}, \tag{13}$$

Then, the set of coupling constraints can be written in the following form

$$\mathbf{x} + \mathbf{B}\boldsymbol{\theta} = \mathbf{0}. \tag{14}$$

Note that $\mathbf{0}$ is a vector of all zeros. Henceforth, let $h(\boldsymbol{\theta}) = \mu N \|\boldsymbol{\theta}\|_2^2$, $\beta = 2\rho$, and $\boldsymbol{\gamma} = \beta\boldsymbol{\lambda}$ the augmented Lagrangian can be equivalently rewritten in the un-scaled ADMM form.

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \ & \sum_{i=1}^{N} \left(\Phi(f_i(\boldsymbol{\theta}_i)) - \mu \|\boldsymbol{\theta}_i\|^2\right) + h(\boldsymbol{\theta}) \\ & + \boldsymbol{\gamma}^T(\mathbf{x} + \mathbf{B}\boldsymbol{\theta}) + \frac{\beta}{2}\|\mathbf{x} + \mathbf{B}\boldsymbol{\theta}\|_2^2. \end{aligned} \tag{15}$$

Note that (15) and the augmented Lagrangian formulated in the main paper (eq (8) in the main paper) are equivalent [2]. Note also that from now on β is sometimes used to refer to ρ in the main paper.

2.1 Monotonicity of the Lagrangian

First, it will be shown that with a sufficiently large β, after each ADMM iteration, the Lagrangian function is non-increasing.

Consider the $(t+1)$-th update cycle of the ADMM iterations. Let $\mathbf{x}^t$, $\boldsymbol{\theta}^t$, and $\boldsymbol{\gamma}^t$ denote the variables carried from the previous iterations and $\mathbf{x}^+$, $\boldsymbol{\theta}^+$, and $\boldsymbol{\gamma}^+$ denote the updated variables, i.e.,$\mathbf{x}^{(t+1)}$, $\boldsymbol{\theta}^{(t+1)}$, and $\boldsymbol{\gamma}^{(t+1)}$, respectively.

As the update steps for the auxiliary variables, which involves minimizing (15) with respect to $\mathbf{x}$, can be solved up to global optimality, the following inequality holds

$$\mathcal{L}_\rho(\mathbf{x}^+, \boldsymbol{\theta}^t, \boldsymbol{\lambda}^t) \leq \mathcal{L}_\rho(\mathbf{x}^t, \boldsymbol{\theta}^t, \boldsymbol{\lambda}^t). \tag{16}$$

After the original variable $\boldsymbol{\theta}$ and the Lagrangian multipliers $\boldsymbol{\gamma}$ are updated, consider the difference between the two Lagrangian functions

$$D_\mathcal{L} = \mathcal{L}_\rho(\mathbf{x}^+, \boldsymbol{\theta}^t, \boldsymbol{\lambda}^t) - \mathcal{L}_\rho(\mathbf{x}^+, \boldsymbol{\theta}^+, \boldsymbol{\lambda}^+). \tag{17}$$

In order to prove the monotonicity of the augmented Lagrangian (15), in the following we prove that with a sufficiently large β, $D_\mathcal{L} \geq 0$.

Since $\boldsymbol{\theta}^+$ minimizes $\mathcal{L}_\rho(\mathbf{x}^+, \boldsymbol{\theta}^t, \boldsymbol{\lambda}^t)$ during the $\boldsymbol{\theta}$ update step, by the optimality condition,

$$\nabla\mathcal{L}_\beta(\boldsymbol{\theta}^+) = \nabla h(\boldsymbol{\theta}^+) + \mathbf{B}^T\boldsymbol{\gamma}^t + \mathbf{B}^T\beta(\mathbf{x}^+ + \mathbf{B}\boldsymbol{\theta}^+) = \mathbf{0}. \tag{18}$$

Note that with the changes of variable, $\boldsymbol{\gamma}$ is updated by

$$\boldsymbol{\gamma}^+ = \boldsymbol{\gamma}^t + \beta(\mathbf{x}^+ + \mathbf{B}\boldsymbol{\theta}^+). \tag{19}$$

Thus, (18) becomes

$$\mathbf{B}^T\boldsymbol{\gamma}^+ = -\nabla h(\boldsymbol{\theta}^+) \tag{20}$$

After some manipulations, $D_{\mathcal{L}}$ can be written as

$$\begin{aligned} D_{\mathcal{L}} = \; & h(\boldsymbol{\theta}^t) - h(\boldsymbol{\theta}^+) + (\boldsymbol{\gamma}^+)^T(\mathbf{B}\boldsymbol{\theta}^t - \mathbf{B}\boldsymbol{\theta}^+) \\ & + \frac{\beta}{2}\|\mathbf{B}\boldsymbol{\theta}^+ - \mathbf{B}\boldsymbol{\theta}\|_2^2 - \frac{1}{\beta}\|\boldsymbol{\gamma}^+ - \boldsymbol{\gamma}^t\|_2^2 \\ = \; & h(\boldsymbol{\theta}^t) - h(\boldsymbol{\theta}^+) + (\mathbf{B}^T\boldsymbol{\gamma}^+)^T(\boldsymbol{\theta}^t - \boldsymbol{\theta}^+) \\ & + \frac{\beta}{2}\|\mathbf{B}\boldsymbol{\theta}^+ - \mathbf{B}\boldsymbol{\theta}\|_2^2 - \frac{1}{\beta}\|\boldsymbol{\gamma}^+ - \boldsymbol{\gamma}^t\|_2^2 \end{aligned} \tag{21}$$

Using (20), (21) becomes

$$\begin{aligned} D_{\mathcal{L}} = \; & h(\boldsymbol{\theta}^t) - h(\boldsymbol{\theta}^+) - \nabla h(\boldsymbol{\theta}^+)^T(\boldsymbol{\theta}^t - \boldsymbol{\theta}^+) \\ & + \frac{\beta}{2}\|\mathbf{B}\boldsymbol{\theta}^+ - \mathbf{B}\boldsymbol{\theta}\|_2^2 - \frac{1}{\beta}\|\boldsymbol{\gamma}^+ - \boldsymbol{\gamma}^t\|_2^2 \end{aligned} \tag{22}$$

Since $\boldsymbol{\gamma}^+ \in \mathrm{Im}(\mathbf{B})$, following [3, Lemma 2],

$$\|\boldsymbol{\gamma}^+ - \boldsymbol{\gamma}\| \leq \|\mathbf{B}^T(\boldsymbol{\gamma}^+ - \boldsymbol{\gamma})\| \leq \|\nabla h(\boldsymbol{\theta}^+) - \nabla h(\boldsymbol{\theta}^t)\| \tag{23}$$

With the definition of $h(\boldsymbol{\theta})$, from (23), it follows that

$$\|\boldsymbol{\gamma}^+ - \boldsymbol{\gamma}\| \leq C\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}^t\|, \tag{24}$$

with $C = 2\mu N$. Thus,

$$-\frac{1}{\beta}\|\boldsymbol{\gamma}^+ - \boldsymbol{\gamma}\|_2^2 \geq -\frac{C^2}{\beta}\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}^t\|_2^2. \tag{25}$$

Furthermore, based on Taylor expansion of the function $h(\boldsymbol{\theta})$,

$$\begin{aligned} h(\boldsymbol{\theta}^t) \geq \; & h(\boldsymbol{\theta}^+) + \nabla h(\boldsymbol{\theta}^+)^T(\boldsymbol{\theta}^t - \boldsymbol{\theta}^+) \\ & + (\boldsymbol{\theta}^t - \boldsymbol{\theta}^+)^T\nabla^2 h(\boldsymbol{\theta}^+)(\boldsymbol{\theta}^t - \boldsymbol{\theta}^+), \end{aligned} \tag{26}$$

which leads to

$$\begin{aligned} & h(\boldsymbol{\theta}^t) - h(\boldsymbol{\theta}^+) - \nabla h(\boldsymbol{\theta}^+)^T(\boldsymbol{\theta}^t - \boldsymbol{\theta}^+) \\ & \geq C\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^+\|_2^2. \end{aligned} \tag{27}$$

By incorporating (22), (25) and (27), it can be deduced that

$$\begin{aligned} D_{\mathcal{L}} \; & \geq \frac{N\beta}{2}\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_2^2 - \frac{C}{\beta}\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}^t\|_2^2 + C\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^+\|_2^2 \\ & \geq \left(\frac{N\beta}{2} - \frac{C^2}{\beta} + C\right)\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^+\|_2^2, \end{aligned} \tag{28}$$

which means if β is sufficiently large, $D_{\mathcal{L}} \geq 0$. In other words, we have proved that, with a sufficiently large β,

$$\mathcal{L}_\rho(\mathbf{x}^+, \boldsymbol{\theta}^+, \boldsymbol{\lambda}^+) \leq \mathcal{L}_\rho(\mathbf{x}^+, \boldsymbol{\theta}^t, \boldsymbol{\lambda}^t). \tag{29}$$

From (16) and (29), it holds that

$$\mathcal{L}_\rho(\mathbf{x}^+, \boldsymbol{\theta}^+, \boldsymbol{\lambda}^+) \leq \mathcal{L}_\rho(\mathbf{x}^t, \boldsymbol{\theta}^t, \boldsymbol{\lambda}^t), \tag{30}$$

or the augmented Lagrangian function (15) is non-increasing for a sufficiently large β.

2.2 Boundedness of the Lagrangian

As our algorithm is a special case of the non-smooth, non-convex optimization problem discussed in [5]. Due to the fact that all the functions are coercive, it can be proved that the Lagrangian is lower bounded for all t and converges as $t \to \infty$ (see [5, Lemma 6.2]). Therefore,

$$\lim_{t\to\infty} \|\boldsymbol{\theta}^+ - \boldsymbol{\theta}^t\|_2 = 0,$$

which, due to (25), leads to

$$\lim_{t\to\infty} \|\boldsymbol{\gamma}^+ - \boldsymbol{\gamma}^t\|_2 = 0.$$

Also, based on the update rule of $\boldsymbol{\gamma}$

$$\lim_{t\to\infty} \|\boldsymbol{\theta}_i^t - \boldsymbol{\theta}^t\|_2 = 0, \ \ \forall i.$$

It then can be concluded that with a sufficiently large β, the ADMM iterations converge to a stationary point $(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$, such that

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_1^* = \boldsymbol{\theta}_2^* = \ldots \boldsymbol{\theta}_N^*.$$

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[3] Qinghua Liu, Xinyue Shen, and Yuantao Gu. Linearized admm for non-convex non-smooth optimization with convergence analysis. *arXiv preprint arXiv:1705.02502*, 2017.

[4] Jaehyun Park and Stephen Boyd. General heuristics for nonconvex quadratically constrained quadratic programming. 2017.

[5] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*, 2015.