Trajectory of Alternating Direction Method of Multipliers and Adaptive Acceleration

Clarice Poon* Jingwei Liang†

Abstract. The alternating direction method of multipliers (ADMM) is one of the most widely used first-order optimisation methods in the literature owing to its simplicity and efficiency. Over the years, different efforts are made to improve the method, such as the inertial technique. By studying the geometric properties of ADMM, we discuss the limitations of current inertial accelerated ADMM and then present and analyse an adaptive acceleration scheme for ADMM. Numerical experiments on problems arising from image processing, statistics and machine learning demonstrate the advantages of the proposed algorithm.

1 Introduction

Consider the following constrained and composite optimisation problem

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} R(x) + J(y) \quad \text{such that} \quad Ax + By = b, \tag{$\mathcal{P}$}$$

where the following basic assumptions are imposed

- (***A*.1**) $R \in \Gamma_0(\mathbb{R}^n)$ and $J \in \Gamma_0(\mathbb{R}^m)$ are proper convex and lower semi-continuous.
- (***A*.2**) $A : \mathbb{R}^n \to \mathbb{R}^p$ and $B : \mathbb{R}^m \to \mathbb{R}^p$ are injective linear operators.
- (***A*.3**) $\mathrm{ri}(\mathrm{dom}(R) \cap \mathrm{dom}(J)) \neq \emptyset$, and the set of minimisers is non-empty.

Over the past years, problem ($\mathcal{P}$) has attracted a great deal of interests as it covers many important problems arising from data science, machine learning, statistics and image processing, etc.; See Section 5 for examples. In the literature, different solvers are proposed to handle the problem, among them the alternating direction method of multipliers (ADMM) is the most prevailing one.

ADMM was first proposed in [20] and becomes increasingly popular recently owing to [10]. The Lagrangian associated to ($\mathcal{P}$) reads

$$\mathcal{L}(x, y; \psi) \stackrel{\text{def}}{=} R(x) + J(y) + \langle \psi,\, Ax + By - b\rangle,$$

and the augmented Lagrangian then simply is:

$$\mathcal{L}_\gamma(x, y; \psi) \stackrel{\text{def}}{=} \mathcal{L}(x, y; \psi) + \tfrac{\gamma}{2}\|Ax + By - b\|^2,$$

where $\gamma > 0$. To find a saddle-point of $\mathcal{L}(x, y; \psi)$, ADMM applies the following iteration

$$\begin{aligned} x_k &= \mathrm{argmin}_{x\in\mathbb{R}^n}\, R(x) + \tfrac{\gamma}{2}\|Ax + By_{k-1} - b + \tfrac{1}{\gamma}\psi_{k-1}\|^2, \\ y_k &= \mathrm{argmin}_{y\in\mathbb{R}^m}\, J(y) + \tfrac{\gamma}{2}\|Ax_k + By - b + \tfrac{1}{\gamma}\psi_{k-1}\|^2, \\ \psi_k &= \psi_{k-1} + \gamma(Ax_k + By_k - b). \end{aligned} \tag{1.1}$$

*Department of Mathematics, University of Bath, Bath, UK. E-mail: cmhsp20@bath.ac.uk.
†DAMTP, University of Cambridge, Cambridge, UK. E-mail: jl993@cam.ac.uk.

Define a new point $z_k \overset{\text{def}}{=} \psi_{k-1} + \gamma A x_k$, then we can rewrite ADMM iteration (1.1) as

$$\begin{aligned} x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax - \tfrac{1}{\gamma}(z_{k-1} - 2\psi_{k-1})\|^2, \\ z_k &= \psi_{k-1} + \gamma A x_k, \\ y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|By + \tfrac{1}{\gamma}(z_k - \gamma b)\|^2, \\ \psi_k &= z_k + \gamma(By_k - b). \end{aligned} \tag{1.2}$$

For the rest of the paper, we will consider the above four-point formulation.

Contributions The contribution of our paper is threefold. First, for the sequence $\{z_k\}_{k\in\mathbb{N}}$ of (1.2), we prove that it has two different types of trajectory:

- When both R, J are non-smooth functions, under the assumption that they are partly smooth (see Definition 2.1), we show that the eventual trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is approximately a spiral which can be characterised precisely if R, J are moreover locally polyhedral.
- When at least one of R, J is smooth, we show that under properly chosen γ, the eventual trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is a straight line.

Then, based on trajectory of $\{z_k\}_{k\in\mathbb{N}}$, we discuss the limitations of the current combination between ADMM and inertial acceleration technique. In Section 3, we distinguish the situations where inertial acceleration will work and when it fails. More precisely: inertial technique will work if the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is or close to a straight line, and will fail if the trajectory is a spiral.

Our core contribution is an adaptive acceleration for ADMM, which is inspired by the trajectory of ADMM and dubbed A^{3}DMM. The limitation of inertial technique, particularly its failure, implies that the right acceleration scheme should be able to follow the trajectory of the sequence. In Section C, we propose an adaptive linear prediction scheme for accelerating ADMM which is able to following the trajectory of the generated sequence. Our proposed A^{3}DMM belongs to the realm of extrapolation method, and provides an alternative interpretation for polynomial extrapolation methods such as Minimal Polynomial Extrapolation (MPE) [13] and Reduced Rank Extrapolation (RRE) [18, 29].

Related works Over the past decades, owing to the tremendous success of inertial acceleration [31, 8], the inertial technique has been widely adapted to accelerate other first-order algorithms. In the realm of ADMM, related work can be found in [32, 22, 19], either from proximal point algorithm perspective or continuous dynamical system. However, to ensure that inertial acceleration works, strong assumptions are imposed on R, J in ($\mathcal{P}$), such as smooth differentiability or strong convexity. When it comes to general non-smooth problems, these works will fail to provide acceleration.

For more generic acceleration techniques, there are extensive works in numerical analysis on the topic of convergence acceleration for sequences. Given an arbitrary sequence $\{z_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ with limit $z^\star$, finding a transformation $\mathcal{E}_k : \{z_{k-j}\}_{j=1}^q \to \bar{z}_k \in \mathbb{R}^n$ such that $\bar{z}_k$ converges faster to $z^\star$. In general, the process by which $\{z_k\}$ is generated is unknown, q is chosen to be a small integer, and $\bar{z}_k$ is referred to as the extrapolation of z_k. Some of the best known examples include Richardson's extrapolation [33], the Δ^2-process of Aitken [2] and Shank's algorithm [35]. We refer to [11, 12, 36] and references therein for a detailed historical perspective on the development of these techniques. Much of the works on the extrapolation of vector sequences was initiated by Wynn [41] who generalized the work of Shank to vector sequences. In the appendix, the formulation of some of these methods are provided. In particular, minimal polynomial extrapolation (MPE) [13] and Reduced Rank Extrapolation (RRE) [18, 29] (which is also a variant of Anderson acceleration developed independently in [4]), which are particularly relevant to this present work (see Section 4.2).

More recently, there has been a series of work on a regularized version of RRE stemming from [34]. We remark however the regularization parameter in these works rely on a grid search based on objective function, their applicability to the general ADMM setting is unclear.

Notations Denote $\mathbb{R}^n$ a n-dimensional Euclidean space equipped with scalar product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. Id denotes the identity operator on $\mathbb{R}^n$. $\Gamma_0(\mathbb{R}^n)$ denotes the class of proper convex and lower-semicontinuous

functions on $\mathbb{R}^n$. For a nonempty convex set $S \subset \mathbb{R}^n$, denote $\mathrm{ri}(S)$ its relative interior, $\mathrm{par}(S)$ the smallest subspace parallel to S and $\mathcal{P}_S$ the projection operator onto S. The sub-differential of a function $R \in \Gamma_0(\mathbb{R}^n)$ is defined by $\partial R(x) \stackrel{\text{def}}{=} \{g \in \mathbb{R}^n | R(x') \geq R(x) + \langle g, x' - x\rangle, \forall x' \in \mathbb{R}^n\}$. The spectral radius of a matrix M is denoted by $\rho(M)$.

2 Trajectory of ADMM

In this section, we discuss the trajectory of the sequence $\{z_k\}_{k\in\mathbb{N}}$ generated by ADMM based on the concept "partial smoothness" which was first introduced in [24].

2.1 Partial smoothness

Let $\mathcal{M} \subset \mathbb{R}^n$ be a C^2-smooth submanifold, denote $\mathcal{T}_{\mathcal{M}}(x)$ the tangent space of $\mathcal{M}$ at a point $x \in \mathcal{M}$.

Definition 2.1 (Partly smooth function [24]). A function $R \in \Gamma_0(\mathbb{R}^n)$ is partly smooth at $\bar{x}$ relative to a set $\mathcal{M}_{\bar{x}}$ if $\partial R(\bar{x}) \neq \emptyset$ and $\mathcal{M}_{\bar{x}}$ is a C^2 manifold around $\bar{x}$, and moreover

Smoothness R restricted to $\mathcal{M}_{\bar{x}}$ is C^2 around $\bar{x}$.

Sharpness The tangent space $\mathcal{T}_{\mathcal{M}_{\bar{x}}}(\bar{x}) = \mathrm{par}(\partial R(\bar{x}))^{\perp}$.

Continuity The set-valued mapping ∂R is continuous at x relative to $\mathcal{M}_{\bar{x}}$.

The class of partly smooth functions at $\bar{x}$ relative to $\mathcal{M}_{\bar{x}}$ is denoted as $\mathrm{PSF}_{\bar{x}}(\mathcal{M}_{\bar{x}})$. Popular examples of partly smooth functions can be found in [25, Chapter 5]. Loosely speaking, a partly smooth function behaves *smoothly* as we move along $\mathcal{M}_{\bar{x}}$, and *sharply* if we move transversal to it.

2.2 Trajectory of ADMM

The iteration of ADMM is non-linear in general owing to the non-smoothness and non-linearity of R and J. However, if they are partly smooth, the local C^2-smoothness allows us to linearise the ADMM iteration, and hence enables us to study the trajectory of sequence generated by the method. We denote $(x^\star, y^\star, \psi^\star)$ a saddle-point of $\mathcal{L}(x, y; \psi)$ and let $z^\star = \psi^\star + \gamma A x^\star$.

To discuss the trajectory of ADMM, we rely on sequence $\{z_k\}_{k\in\mathbb{N}}$. Define $v_k \stackrel{\text{def}}{=} z_k - z_{k-1}$ and $\theta_k \stackrel{\text{def}}{=} \arccos(\frac{\langle v_k, v_{k-1}\rangle}{\|v_k\|\|v_{k-1}\|})$ the angle between v_k, v_{k-1}. We use $\{\theta_k\}_{k\in\mathbb{N}}$ to characterise the trajectory of $\{z_k\}_{k\in\mathbb{N}}$. Given $(x^\star, y^\star, \psi^\star)$, the first-order optimality condition entails $-A^T\psi^\star \in \partial R(x^\star)$ and $-B^T\psi^\star \in \partial J(y^\star)$, below we impose

$$-A^T\psi^\star \in \mathrm{ri}\big(\partial R(x^\star)\big) \quad \text{and} \quad -B^T\psi^\star \in \mathrm{ri}\big(\partial J(y^\star)\big). \tag{ND}$$

Both R, J are non-smooth Suppose $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}^R_{x^\star}), J \in \mathrm{PSF}_{y^\star}(\mathcal{M}^J_{y^\star})$ are partly smooth, denote $T^R_{x^\star}, T^J_{y^\star}$ the tangent spaces of $\mathcal{M}^R_{x^\star}, \mathcal{M}^J_{y^\star}$ at $x^\star, y^\star$. Let $A_R \stackrel{\text{def}}{=} A \circ \mathcal{P}_{T^R_{x^\star}}, B_J \stackrel{\text{def}}{=} B \circ \mathcal{P}_{T^J_{y^\star}}$ and T_{A_R}, T_{B_J} be the range of A_R, B_J respectively. Denote $(\alpha_j)_{j=1,\ldots}$ the Principal angles (see Section A.2 in the appendix for definition) between T_{A_R}, T_{B_J}, and let α_F, α' be the smallest and 2nd smallest of α_j which are yet larger than 0.

Theorem 2.2. *For problem ($\mathcal{P}$) and ADMM iteration (1.1), assume that conditions (**A.1**)-(**A.3**) are true, then (x_k, y_k, ψ_k) converges to a saddle point $(x^\star, y^\star, \psi^\star)$ of $\mathcal{L}(x, y; \psi)$. Suppose that $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}^R_{x^\star}), J \in \mathrm{PSF}_{y^\star}(\mathcal{M}^J_{y^\star})$ and condition (ND) holds, then*

(i) *There exists a matrix M_{ADMM} such that $v_k = M_{\mathrm{ADMM}} v_{k-1} + o(\|v_{k-1}\|)$ holds for all k large enough.*

(ii) *If moreover, R, J are locally polyhedral around $x^\star, y^\star$, then $v_k = M_{\mathrm{ADMM}} v_{k-1}$ with M_{ADMM} being normal and having eigenvalues of the form $\cos(\alpha_j)e^{\pm \mathrm{i}\alpha_j}$, and $\cos(\theta_k) = \cos(\alpha_F) + O(\eta^{2k})$ with $\eta = \cos(\alpha')/\cos(\alpha_F)$.*

Remark 2.3. The result indicates that, when both R, J are locally polyhedral, the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is a spiral. For the case R, J being general partly smooth function, though we cannot prove, numerical evidence shows that the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ could be either straight line or also a spiral.

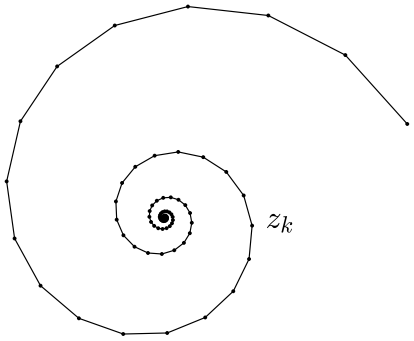
R or/and J is smooth Now we consider the case that at least one function out of R, J is smooth. For simplicity, consider that R is smooth and J remains non-smooth.

Proposition 2.4. *For problem ($\mathcal{P}$) and ADMM iteration (1.1), assume that conditions ($\mathcal{A}$.1)-($\mathcal{A}$.3) are true, then (x_k, y_k, ψ_k) converges to a saddle point $(x^\star, y^\star, \psi^\star)$ of $\mathcal{L}(x, y; \psi)$. Suppose R is locally C^2 around $x^\star$, $J \in \mathrm{PSF}_{y^\star}(\mathcal{M}^J_{y^\star})$ is partly smooth and condition (ND) holds for J, then Theorem 2.2(i) holds for all k large enough. If moreover, A is full rank square matrix, then all the eigenvalues of M_{ADMM} are real for $\gamma > \|(A^TA)^{-\frac{1}{2}}\nabla^2 R(x^\star)(A^TA)^{-\frac{1}{2}}\|$.*

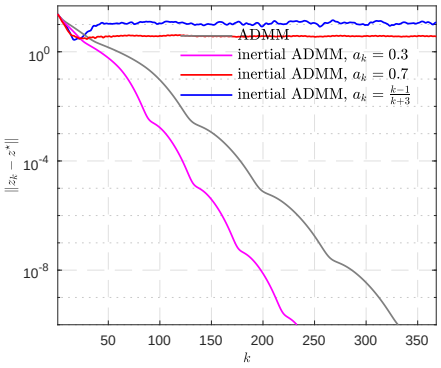
Remark 2.5. The real spectrum of M, numerical evidence shows that the eventual trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is a straight line, which is different from the case where both functions are non-smooth. If $o(\|v_{k-1}\|)$ is vanishing fast enough, we can also prove that $\theta_k \to 0$.

When $\gamma \leq \|(A^TA)^{-\frac{1}{2}}\nabla^2 R(x^\star)(A^TA)^{-\frac{1}{2}}\|$, M_{ADMM} will have complex eigenvalues, however the trajectory could be either spiral or straight line depending the leading eigenvalue. If both R, J are smooth, M will also have real spectrum under proper choice of γ.

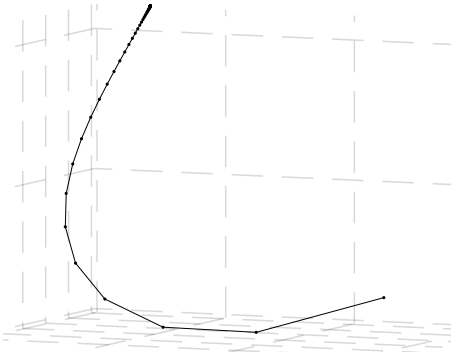
In Figure 1 (a) and (c), we present two examples of the trajectory of ADMM. Subfigure (a) shows a spiral trajectory in $\mathbb{R}^2$ which is obtained from solving a polyhedral problem, while subfigure (c) is an eventual straight line trajectory in $\mathbb{R}^3$.



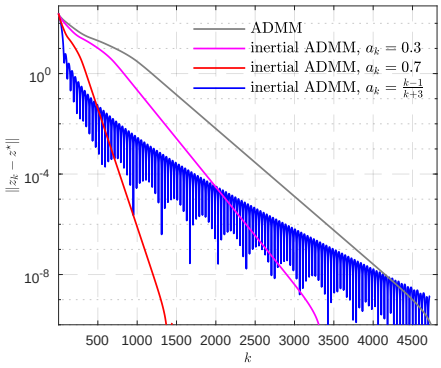
(a) Spiral



(b) $\gamma = 1$



(c) Eventual straight line



(d) $\gamma = \|K\|^2$

Figure 1: Trajectory of sequence $\{z_k\}_{k\in\mathbb{N}}$ and effects of inertial on ADMM. (a) Spiral trajectory of ADMM; (b) failure of inertial ADMM on spiral trajectory; (c) Eventual straight line trajectory; (d) success of inertial ADMM on straight line trajectory.

3 The failure of inertial acceleration

We use the LASSO problem as an example to demonstrate the effects of applying the inertial technique to ADMM, especially when it failures. One simple approach for combining inertial technique with ADMM is described below

$$\begin{aligned} x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax - \tfrac{1}{\gamma}(\bar{z}_{k-1} - 2\psi_{k-1})\|^2, \\ z_k &= \psi_{k-1} + \gamma A x_k, \\ \bar{z}_k &= z_k + a_k(z_k - z_{k-1}), \\ y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|By + \tfrac{1}{\gamma}(\bar{z}_k - \gamma b)\|^2, \\ \psi_k &= \bar{z}_k + \gamma(By_k - b), \end{aligned} \tag{3.1}$$

which considers only the momentum of $\{z_k\}_{k\in\mathbb{N}}$ without any stronger assumptions on R, J. The above scheme can reformulated as an instance of inertial Proximal Point Algorithm, guaranteed to be convergent for $a_k < \frac{1}{3}$ [3]; We refer to [32] or [25, Chapter 4.3] for more details.

The formulation of LASSO in the form of ($\mathcal{P}$) reads

$$\min_{x,y\in\mathbb{R}^n} \mu\|x\|_1 + \tfrac{1}{2}\|Ky - f\|^2 \quad \text{such that} \quad x - y = 0, \tag{3.2}$$

where $K \in \mathbb{R}^{m\times n}$, $m < n$ is a random Gaussian matrix. Since $\frac{1}{2}\|Ky - f\|^2$ is quadratic, owing to Proposition 2.4, the eventual trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is a straight line if $\gamma > \|K\|^2$, and a spiral for some $\gamma \leq \|K\|^2$. Therefore, we consider two different choices of γ which are $\gamma = 1$ and $\gamma = \|K\|^2$, and for each γ, four different choices of a_k are considered

$$a_k \equiv 0.3, \quad a_k \equiv 0.7 \quad \text{and} \quad a_k = \tfrac{k-1}{k+3}.$$

The 3rd choice of a_k corresponds to FISTA [14]. Numerical results are shown in Figure 1 (b) and (d),

- When $\gamma = 1$, the inertial scheme works only for $a_k \equiv 0.3$, which is due to that fact that the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is a spiral for $\gamma = 1$. As a result, the direction $z_k - z_{k-1}$ is not pointing towards $z^\star$, hence unable to provide satisfactory acceleration.
- When $\gamma = \|K\|^2$, all choices of a_k work since $\{z_k\}_{k\in\mathbb{N}}$ eventually forms a straight line. Among these four choices of a_k, $a_k \equiv 0.7$ is the fastest, while $a_k = \frac{k-1}{k+3}$ eventually is the slowest.

It should be noted that, though ADMM is faster for $\gamma = 1$ than $\gamma = \|K\|^2$, our main focus here is to demonstrate how the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ affects the outcome of inertial acceleration.

The above comparisons, particularly for $\gamma = 1$ implies that the trajectory of the sequence $\{z_k\}_{k\in\mathbb{N}}$ is crucial for the acceleration outcome of the inertial scheme. Since the trajectories of ADMM depends on the properties of R, J and choice of γ, this implies that the right scheme that can achieve uniform acceleration despite R, J and γ should be able to adapt itself to the trajectory of the method.

4 A^{3}DMM: adaptive acceleration for ADMM

The previous section shows the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ eventually settles onto a regular path *i.e.* either straight line or spiral. In this section, we exploit this regularity to design adaptive acceleration for ADMM, which is called "A^{3}DMM"; See Algorithm 1.

The update of $\bar{z}_k$ in (3.1) can be viewed as a special case of the following extrapolation

$$\bar{z}_k = \mathcal{E}(z_k, z_{k-1}, \cdots, z_{k-q}), \tag{4.1}$$

for the choice of $q = 1$. The idea is: given $\{z_{k-j}\}_{j=0}^{q+1}$, define $v_j \overset{\text{def}}{=} z_j - z_{j-1}$ and predict the future iterates by considering how the past directions $v_{k-1}, \ldots, v_{k-q}$ approximate the latest direction v_k. In particular, define $V_{k-1} \overset{\text{def}}{=} \begin{bmatrix} v_{k-1}, \cdots, v_{k-q} \end{bmatrix} \in \mathbb{R}^{n\times q}$, and let $c_k \overset{\text{def}}{=} \operatorname{argmin}_{c\in\mathbb{R}^q}\|V_{k-1}c - v_k\|^2 = \|\sum_{j=1}^q c_j v_{k-j} - v_k\|^2$. The idea is then that $V_k c_k \approx v_{k+1}$ and so, $\bar{z}_{k,1} \overset{\text{def}}{=} z_k + V_k c \approx z_{k+1}$. By iterating this s times, we obtain $\bar{z}_{k,s} \approx z_{k+s}$.

More precisely, given $c \in \mathbb{R}^q$, define the mapping H by $H(c) = \left[\begin{array}{c|c} c_{1:q-1} & \mathrm{Id}_{q-1} \\ c_q & 0_{1,q-1} \end{array}\right] \in \mathbb{R}^{q\times q}$. Let $C_k = H(c_k)$, note that $V_k = V_{k-1}C_k$. Define $\bar{V}_{k,0} \overset{\text{def}}{=} V_k$ and for $s \geq 1$, define

$$\bar{V}_{k,s} \overset{\text{def}}{=} \bar{V}_{k,s-1}C_k \overset{\text{def}}{=} V_k C_k^s,$$

where C_k^s is the power of C_k. Let $(C)_{(:,1)}$ be the first column of matrix C, then

$$\bar{z}_{k,s} = z_k + \sum_{i=1}^s (\bar{V}_{k,i})_{(:,1)} = z_k + \sum_{i=1}^s V_k (C_k^i)_{(:,1)} = z_k + V_k \Big(\sum_{i=1}^s C_k^i\Big)_{(:,1)}, \tag{4.2}$$

which is the desired trajectory following extrapolation scheme. Now define the extrapolation parameterised by s, q as

$$\mathcal{E}_{s,q}(z_k, \cdots, z_{k-q-1}) \overset{\text{def}}{=} V_k \Big(\sum_{i=1}^s C_k^i\Big)_{(:,1)},$$

we obtain the following trajectory following adaptive acceleration for ADMM.

Algorithm 1: A^{3}DMM: Adaptive Acceleration for ADMM

Initial: Let $s \geq 1, q \geq 1$ be integers and $p = q + 1$. Let $\bar{z}_0 = z_0 \in \mathbb{R}^n$ and $V_0 = 0 \in \mathbb{R}^{n\times q}$.

Repeat:

- For $k \geq 1$:

$$\begin{aligned} y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|By + \tfrac{1}{\gamma}(\bar{z}_{k-1} - \gamma b)\|^2, \\ \psi_k &= \bar{z}_{k-1} + \gamma(By_k - b), \\ x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax - \tfrac{1}{\gamma}(\bar{z}_{k-1} - 2\psi_k)\|^2, \\ z_k &= \psi_k + \gamma A x_k, \\ v_k &= z_k - z_{k-1} \quad \text{and} \quad V_k = [v_k, V_{k-1}(:, 1:q-1)]. \end{aligned}$$

- If $\mathrm{mod}(k,p) = 0$: Compute C_k as described above, if $\rho(C_k) < 1$:

$$\bar{z}_k = z_k + a_k \mathcal{E}_{s,q}(z_k, \cdots, z_{k-q-1}).$$

- If $\mathrm{mod}(k,p) \neq 0$: $\bar{z}_k = z_k$.

Until: $\|v_k\| \leq \mathrm{tol}$.

Remark 4.1.

- When $\mathrm{mod}(k,p) \neq 0$, one can also consider $\bar{z}_k = z_k + a_k(z_k - z_{k-1})$ with properly chosen a_k.
- A^{3}DMM carries out p standard ADMM iterations to set up the extrapolation step $\mathcal{E}_{s,q}$. As $\mathcal{E}_{s,q}$ contains the sum of the powers of C_k which is guaranteed to be convergent when $\rho(C_k) < 1$. Therefore, we only apply $\mathcal{E}_{s,q}$ when the spectral radius $\rho(C_k) < 1$ is true. In this case, there is a closed form expression for $\mathcal{E}_{s,q}$ when $s = +\infty$; See Eq. (4.4).
- The purpose of adding a_k in front of $\mathcal{E}_{s,q}(z_k, \cdots, z_{k-q-1})$ is so that we can control the value of a_k to ensure the convergence of the algorithm; See below the discussion.

4.1 Convergence of A^{3}DMM

To discuss the convergence of A^{3}DMM, we shall treat the algorithm as a perturbation of the original ADMM. If the perturbation error is absolutely summable, then we obtain the convergence of A^{3}DMM. More precisely, let $\varepsilon_k \in \mathbb{R}^n$ whose value takes

$$\varepsilon_k = \begin{cases} 0 : \mathrm{mod}(k,p) \neq 0 \ \text{ or } \ \mathrm{mod}(k,p) = 0 \ \&\ \rho(C_k) \geq 1, \\ a_k \mathcal{E}_{s,q}(z_k, \cdots, z_{k-q-1}) : \mathrm{mod}(k,p) = 0 \ \&\ \rho(C_k) < 1. \end{cases}$$

Suppose the fixed-point formulation of ADMM can be written as $z_k = \mathcal{F}(z_{k-1})$ for some $\mathcal{F}$ (see Section B of the appendix for details). Then Algorithm 1 can be written as

$$z_k = \mathcal{F}(z_{k-1} + \varepsilon_{k-1}). \tag{4.3}$$

Owing to (4.3), we can obtain the following convergence for Algorithm 1 which is based on the classic convergence result of inexact Krasnosel'skiĭ-Mann fixed-point iteration [5, Proposition 5.34].

Proposition 4.2. *For problem ($\mathcal{P}$) and Algorithm 1, suppose that the conditions ($\mathcal{A}$.1)-($\mathcal{A}$.3) are true. If moreover, $\sum_k \|\varepsilon_k\| < +\infty$, $z_k \to z^\star \in \mathrm{fix}(\mathcal{F}) \overset{\text{def}}{=} \{z \in \mathbb{R}^p : z = \mathcal{F}(z)\}$ and (x_k, y_k, ψ_k) converges to $(x^\star, y^\star, \psi^\star)$ which is a saddle point of $\mathcal{L}(x, y; \psi)$.*

On-line updating rule The summability condition $\sum_k \|\varepsilon_k\| < +\infty$ in general cannot be guaranteed. However, it can be enforced by a simple online updating rule. Let $a \in [0, 1]$ and $b, \delta > 0$, then a_k can be determined by $a_k = \min\{a, b/(k^{1+\delta}\|z_k - z_{k-1}\|)\}$.

Inexact A^{3}DMM Observe that in A^{3}DMM, when A, B are non-trivial, in general there are no closed form solutions for x_k and y_k. Take x_k for example, suppose it is computed approximately, then in z_k there will be another approximation error ε'_k, and consequently

$$z_k = \mathcal{F}(z_{k-1} + \varepsilon_{k-1} + \gamma\varepsilon'_{k-1}).$$

If there holds $\sum_k \|\varepsilon'_{k-1}\| < +\infty$, Proposition 4.2 remains true for the above perturbation form.

4.2 Acceleration guarantee for A^{3}DMM

We have so far alluded to the idea that the extrapolated point $\bar{z}_{k,s}$ defined in (4.2) (which depends only on $\{z_{k-j}\}_{j=0}^q$) is an approximation to z_{k+s}. In this section, we make precise this statement.

Relationship to MPE and RRE We first show that $\bar{z}_{k,\infty}$ is (almost) equivalent to MPE. Recall that given a square matrix C, if its Neumann series is convergent, then there holds $(\mathrm{Id} - C)^{-1} = \sum_{i=0}^{+\infty} C^i$. Now for the summation of the power of C_k in (4.2), when $s = +\infty$, we have

$$\sum\nolimits_{i=1}^{+\infty} C_k^i = C_k \sum\nolimits_{i=0}^{+\infty} C_k^i = C_k(\mathrm{Id} - C_k)^{-1} = (\mathrm{Id} - C_k)^{-1} - \mathrm{Id}.$$

Back to (4.2), then we get

$$\begin{aligned} \bar{z}_{k,\infty} &\overset{\text{def}}{=} z_k + V_k\big((\mathrm{Id} - C_k)^{-1} - \mathrm{Id}\big)_{(:,1)} = z_k - v_k + V_k\big((\mathrm{Id} - C_k)^{-1}\big)_{(:,1)} \\ &= z_{k-1} + V_k\big((\mathrm{Id} - C_k)^{-1}\big)_{(:,1)} = \frac{1}{1 - \sum_{i=1}^s c_{k,i}}\big(z_k - \sum\nolimits_{j=1}^{q-1} c_{k,j} z_{k-j}\big), \end{aligned} \tag{4.4}$$

which turns out to be MPE, with the slight difference of taking the weighted sum of $\{z_j\}_{j=k-q+1}^k$ as opposed to the weighted sum of $\{z_j\}_{j=k-q}^{k-1}$ (See appendix for more details of MPE). Note that if the coefficients c is computed in the following way: $b \in \mathrm{argmin}_{a \in \mathbb{R}^{q+1}, \sum_j a_j = 1} \|\sum_{j=0}^q a_j v_{k-j}\|$ and $b_0 \neq 0$ and define $c_j \overset{\text{def}}{=} -b_j/b_0$ for $j = 1, \ldots, q$. Then,

$$\big(1 - \sum\nolimits_{i=1}^q c_i\big)^{-1} = \frac{b_0}{b_0 + \sum_{j=1}^q b_j} = b_0,$$

and $\bar{z}_{k,\infty} = \sum_{j=0}^{q-1} b_j z_{k-j}$ is precisely the RRE update (again with the slight difference of summing over iterates shifted by one iteration).

Acceleration guarantee for A^{3}DMM Let $\{z_k\}_{k \in \mathbb{N}}$ be a sequence in $\mathbb{R}^n$ and let $v_k \overset{\text{def}}{=} z_k - z_{k-1}$. Assume that $v_k = M v_{k-1}$ for some $M \in \mathbb{R}^{n \times n}$. Denote $\lambda(M)$ the spectrum of M. The following proposition provides control on the extrapolation error for $\bar{z}_{k,s}$ from (4.2).

Proposition 4.3. *Define the coefficient fitting error by $\epsilon_k \overset{\text{def}}{=} \min_{c \in \mathbb{R}^q} \|V_{k-1}c - v_k\|$.*

(i) *For $s \in \mathbb{N}$, we have*

$$\|\bar{z}_{k,s} - z^\star\| \leq \|z_{k+s} - z^\star\| + B_s \epsilon_k. \tag{4.5}$$

where $B_s \overset{\text{def}}{=} \sum_{\ell=1}^s \|M^\ell\| |\sum_{i=0}^{s-\ell} (C_k^i)_{(1,1)}|$. If $\rho(M) < 1$ and $\rho(C_k) < 1$, then $\sum_i c_{k,i} \neq 1$ and B_s is uniformly bounded in s. For $s = +\infty$, $B_\infty \overset{\text{def}}{=} |1 - \sum_i c_{k,i}|^{-1} \sum_{\ell=1}^\infty \|M\|^\ell$

(ii) *Suppose that M is diagonalisable. Let $(\lambda_j)_j$ denote its distinct eigenvalues ordered such that $|\lambda_j| \geq |\lambda_{j+1}|$ and $|\lambda_1| = \rho(M) < 1$. Suppose that $|\lambda_q| > |\lambda_{q+1}|$.*

- *Asymptotic bound (fixed q and as $k \to +\infty$): $\epsilon_k = \mathcal{O}(|\lambda_{q+1}|^k)$.*
- *Nonasymptotic bound (fixed q and k): Suppose that $\lambda(M)$ is real-valued and contained in the interval $[\alpha, \beta]$ with $-1 < \alpha < \beta < 1$. Then,*

$$\frac{\epsilon_k}{1 - \sum_i c_{k,i}} \leq K\beta^{k-q}\left(\frac{\sqrt{\eta}-1}{\sqrt{\eta}+1}\right)^q \tag{4.6}$$

where $K \stackrel{\text{def}}{=} 2\|z_0 - z^\star\|\|(\mathrm{Id} - M)^{\frac{1}{2}}\|$ and $\eta = \frac{1-\alpha}{1-\beta}$.

Remark 4.4.

- From Theorem 2.2(ii), when R and J are both polyhedral, we have a perfect local linearisation with the corresponding linearisation matrix being normal and hence, the conditions of Proposition 4.3 holds for all k large enough. The first bound (i) shows that the extrapolated point $\bar{z}_{k,s}$ moves along the true trajectory as s increases, up to the fitting error ϵ_k. Although $\bar{z}_{k,\infty}$ is essentially an MPE update which is known to satisfy error bound (4.6) (see [37]), this proposition offers a further interpretation of these extrapolation methods in terms of following the "sequence trajectory", and combined with our local analysis of ADMM, provides justification of these methods for the acceleration of non-smooth optimisation problems.
- Proposition 4.3 (ii) shows that extrapolation improves the convergence rate from $\mathcal{O}(|\lambda_1|^k)$ to $\mathcal{O}(|\lambda_{q+1}|^k)$, and the nonasymptotic bound shows that the improvement of extrapolation is optimal in the sense of Nesterov [31]. Recalling the form of the eigenvalues of M from Theorem 2.2, in the case of two nonsmooth polyhedral terms, we must have $|\lambda_{2j-1}| = |\lambda_{2j}| > |\lambda_{2j+1}|$ for all $j \geq 1$. Hence, no acceleartion can be guaranteed or observed when $q = 1$, while the choice of $q = 2$ provides guaranteed acceleration.

5 Numerical experiments

Below we present numerical experiments on affine constrained minimisation (*e.g.* Basis Pursuit), LASSO, quadratic programming and image processing problems to demonstrate the performance of the proposed scheme. In the numerical comparison below, we mainly compare with the original ADMM and its inertial version (3.1) with fixed $a_k \equiv 0.3$. For the proposed A^{3}DMM, two settings are considered: $(q, p, s) = (6, 7, 100)$ and $(q, p, s) = (6, 7, +\infty)$. The quantity we compare is $\|x_k - x^\star\|$.

5.1 Affine constrained minimisation

Consider the following constrained problem

$$\min_{x \in \mathbb{R}^n} R(x) \quad \text{such that} \quad Kx = f. \tag{5.1}$$

Denote the set $\Omega \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : Kx = f\}$, and ι_Ω its indicator function. Then (5.1) can be written as

$$\min_{x,y \in \mathbb{R}^n} R(x) + \iota_\Omega(y) \quad \text{such that} \quad x - y = 0, \tag{5.2}$$

which is special case of ($\mathcal{P}$) with $A = \mathrm{Id}$, $B = -\mathrm{Id}$ and $b = 0$. Here K is generated from the standard Gaussian ensemble, and the following three choices of R are considered:

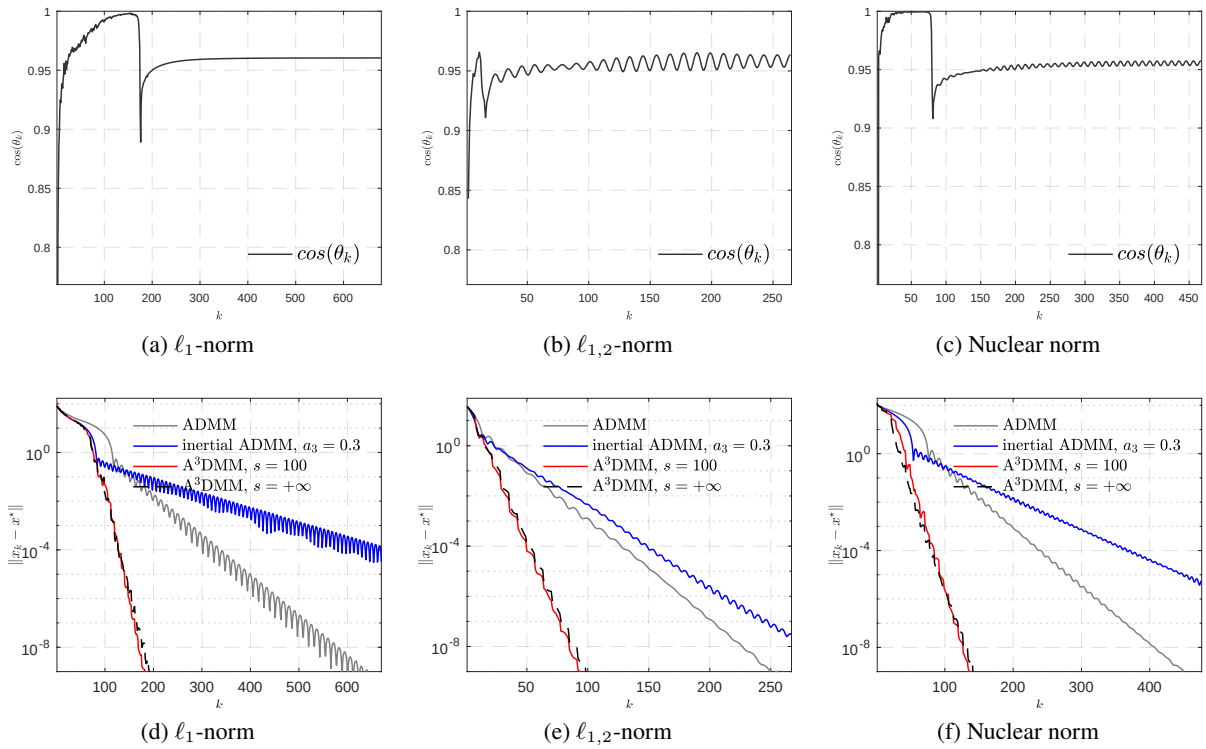
ℓ_1-norm $(m, n) = (512, 2048)$, solution $x^\star$ is 128-sparse;
$\ell_{1,2}$-norm $(m, n) = (512, 2048)$, solution $x^\star$ has 32 non-zero blocks of size 4;
Nuclear norm $(m, n) = (1448, 4096)$, solution $x^\star$ has rank of 4.

The property of $\{\theta_k\}_{k\in\mathbb{N}}$ is shown in Figure 2 (a)-(c). Note that the indicator function $\iota_\Omega(y)$ in (5.2) is polyhedral since Ω is an affine subspace,

- As ℓ_1-norm is polyhedral, we have in Figure 2(a) that θ_k is converging to a constant which complies with Theorem 2.2(ii).
- Since $\ell_{1,2}$-norm and nuclear norm are no longer polyhedral functions, we have that θ_k eventually oscillates in a range, meaning that the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is an elliptical spiral.

Comparisons of the four schemes are shown below in Figure 2 (d)-(f):

- Since both functions in (5.2) are non-smooth, the eventual trajectory of $\{z_k\}_{k\in\mathbb{N}}$ for ADMM is spiral. Inertial ADMM fails to provide acceleration locally.
- A^{3}DMM is faster than both ADMM and inertial ADMM. For the two different settings of A^{3}DMM, their performances are very close.



(a) ℓ_1-norm (b) $\ell_{1,2}$-norm (c) Nuclear norm

(d) ℓ_1-norm (e) $\ell_{1,2}$-norm (f) Nuclear norm

Figure 2: Performance comparisons and $\{\theta_k\}_{k\in\mathbb{N}}$ of ADMM for affine constrained problem.

5.2 LASSO

We consider again the LASSO problem (3.2) with three datasets from LIBSVM[1]. The numerical experiments are provided below in Figure 3.

It can be observed that the proposed A^{3}DMM is significantly faster than the other schemes, especially for $s = +\infty$. Between ADMM and inertial ADMM, different from the previous example, the inertial technique can provided consistent acceleration for all three examples.

5.3 Quadratic programming

Consider the following quadratic optimisation problem

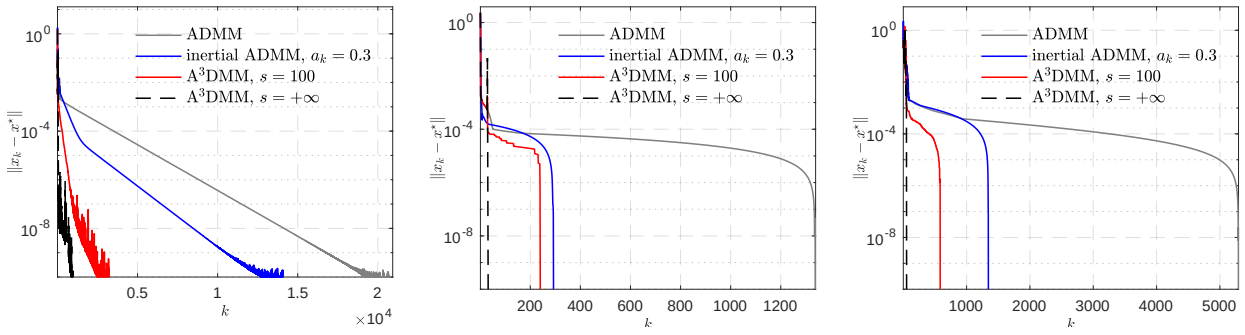
$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \tfrac{1}{2}x^TQx + \langle q,\, x\rangle, \\ \text{such that} \quad & x_i \in [\ell_i, r_i], \quad i = 1, ..., n. \end{aligned} \tag{5.3}$$

Define the constraint set $\Omega = \{x \in \mathbb{R}^n : x_i \in [\ell_i, r_i],\ \ i = 1, ..., n\}$, then (5.3) can be written as

$$\min_{x,y\in\mathbb{R}^n} \quad \tfrac{1}{2}x^TQx + \langle q,\, x\rangle + \iota_\Omega(y) \quad \text{such that} \quad x - y = 0,$$

which is special case of ($\mathcal{P}$) with $A = \mathrm{Id}$, $B = -\mathrm{Id}$ and $b = 0$.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

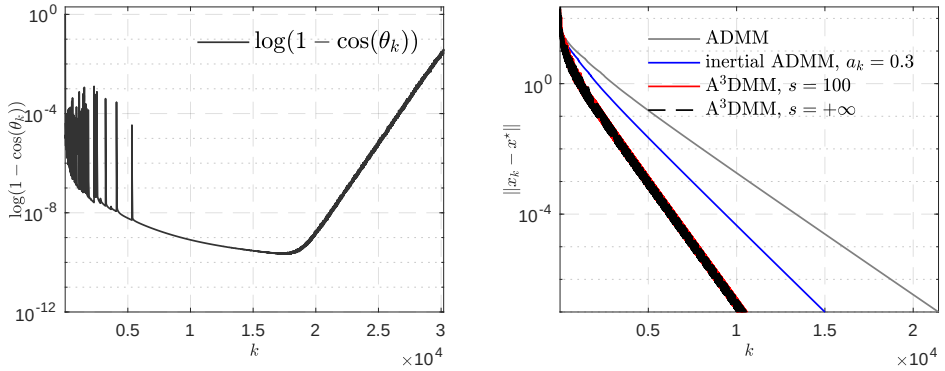


(a) `covtype` (b) `ijcnn1` (c) `phishing`

Figure 3: Performance comparisons for LASSO problem.

The angle θ_k of ADMM and the performances of the four schemes are provided in Figure (4), from which we observed that

- The angle θ_k is decreasing to 0 at the beginning and then starts to increasing for $k \geq 2 \times 10^4$. This is mainly due to the fact that for $k \geq 2 \times 10^4$, the effects of machine error is becoming increasingly larger.
- Consistent with the observations in Section 5, the proposed A^{3}DMM schemes provides the best performance.



(a) Angle θ_k (b) Comparison of $\|x_k - x^\star\|$

Figure 4: Performance comparisons and $\{\theta_k\}_{k\in\mathbb{N}}$ of ADMM for quadratic programming.

5.4 Total variation based image inpainting

Now we consider a total variation (TV) based image inpainting problem. Let $u \in \mathbb{R}^{n\times n}$ be an image and $\mathcal{S} \in \mathbb{R}^{n\times n}$ be a Bernoulli matrix, the observation of u under $\mathcal{S}$ is $f = \mathcal{P}_{\mathcal{S}}(u)$. The TV based image inpainting can be formulated as

$$\min_{x\in\mathbb{R}^{n\times n}} \|\nabla x\|_1 \quad \text{such that} \quad \mathcal{P}_{\mathcal{S}}(x) = f. \tag{5.4}$$

Define $\Omega \overset{\text{def}}{=} \{x \in \mathbb{R}^{n\times n} : \mathcal{P}_{\mathcal{S}}(x) = f\}$, then (5.4) becomes

$$\min_{x\in\mathbb{R}^{n\times n}} \|y\|_1 + \iota_\Omega(x) \quad \text{such that} \quad \nabla x - y = 0, \tag{5.5}$$

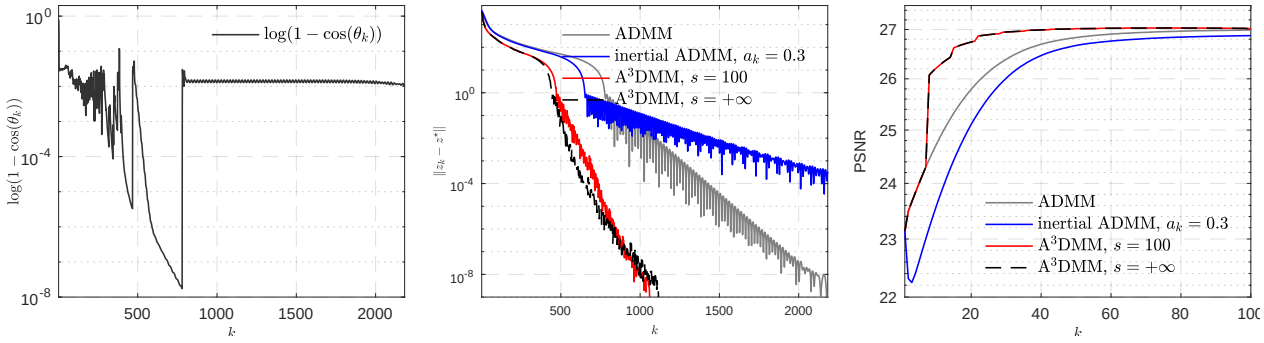
which is special case of ($\mathcal{P}$) with $A = \nabla$, $B = -\mathrm{Id}$ and $b = 0$. For the update of x_k, we have from (1.2) that

$$x_k = \operatorname{argmin}_{x\in\mathbb{R}^{n\times n}} \iota_\Omega(x) + \tfrac{\gamma}{2}\|\nabla x - \tfrac{1}{\gamma}(\bar{z}_{k-1} - 2\psi_{k-1})\|^2,$$

which does not admit closed form solution. In the implementation, finite-step FISTA is applied to roughly solve the above problem.

In the experiment, the `cameraman` image is used, and 50% of the pixels is removed randomly. The angle θ_k of ADMM and the comparisons of the four schemes are provided in Figure 5:

- Though both functions in (5.5) are polyhedral, since the subproblem of x_k is solved approximately, the eventual angle actually is oscillating instead of being a constant.
- Inertial ADMM again is slower than the original ADMM as the trajectory of ADMM is a spiral.
- For the two A^{3}DMM schemes, their performances are close as previous examples.
- For PSNR the image quality assessment, Figure 5(c) implies that A^{3}DMM is also the best.



(a) Angle $\{\theta_k\}_{k\in\mathbb{N}}$ of ADMM (b) Comparison of $\|x_k - x^\star\|$ (c) PSNR value

Figure 5: Property of $\{\theta_k\}_{k\in\mathbb{N}}$, performance comparison and image quality of ADMM for TV based image inpainting.

We also compare the visual quality of the images obtained by the four schemes for the 8'th iteration, which is shown below in Figure 6. Since we choose $(q, p) = (6, 7)$, for $k = 8$ both A^{3}DMM applies only one step adaptive acceleration step, and the image quality (2nd row of Figure 6) is much better than the 1st row of ADMM and inertial ADMM.

6 Conclusions

In this article, by analysing the trajectory of the fixed point sequences associated to ADMM and extrapolating along the trajectory, we provide an alternative derivation of these methods. Furthermore, our local linear analysis allows for the application of previous results on extrapolation methods, and hence provides guaranteed (local) acceleration. Extension of the proposed acceleration framework to general first-order methods is ongoing.

References

[1] P-A. Absil, R. Mahony, and J. Trumpf. An extrinsic look at the Riemannian Hessian. In *Geometric Science of Information*, pages 361–368. Springer, 2013.

[2] A. C. Aitken. Xxv.–on Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, 46:289–305, 1927.

[3] F. Alvarez and H. Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1-2):3–11, 2001.

[4] D. G. Anderson. Iterative procedures for nonlinear integral equations. *J. ACM*, 12(4):547–560, October 1965.

[5] H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

(a) ADMM (b) Inertial ADMM

(c) A^{3}DMM, $s = 100$ (d) A^{3}DMM, $s = +\infty$

Figure 6: Comparison of image quality at the 8'th iteration of ADMM, inertial ADMM and two A^{3}DMM schemes.

[6] H. H. Bauschke, J. Y. Bello Cruz, T. T. A. Nghia, H. M. Pha, and X. Wang. Optimal rates of linear convergence of relaxed alternating projections and generalized Douglas–Rachford methods for two subspaces. *Numerical Algorithms*, 73(1):33–76, 2016.

[7] H. H. Bauschke, JY B. Cruz, T. TA Nghia, H. M. Phan, and X. Wang. The rate of linear convergence of the douglas–rachford algorithm for subspaces is the cosine of the friedrichs angle. *Journal of Approximation Theory*, 185:63–79, 2014.

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[9] P Borwein, Christopher Pinner, and I Pritsker. Monic integer chebyshev problem. *Mathematics of computation*, 72(244):1901–1916, 2003.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[11] C. Brezinski. Convergence acceleration during the 20th century. *Numerical Analysis: Historical Developments in the 20th Century*, page 113, 2001.

[12] C. Brezinski and M. R. Zaglia. *Extrapolation methods: theory and practice*, volume 2. Elsevier, 2013.

[13] S. Cabay and L. W. Jackson. A polynomial extrapolation method for finding limits and antilimits of vector sequences. *SIAM Journal on Numerical Analysis*, 13(5):734–752, 1976.

[14] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

[15] I. Chavel. *Riemannian geometry: a modern introduction*, volume 98. Cambridge University Press, 2006.

[16] L. Demanet and X. Zhang. Eventual linear convergence of the douglas-rachford iteration for basis pursuit. *Mathematics of Computation*, 85(297):209–238, 2016.

[17] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439, 1956.

[18] R. P. Eddy. Extrapolating to the limit of a vector sequence. In *Information linkage between applied mathematics and industry*, pages 387–396. Elsevier, 1979.

[19] G. França, D. P. Robinson, and R. Vidal. Admm and accelerated admm as continuous dynamical systems. *arXiv preprint arXiv:1805.06579*, 2018.

[20] D. Gabay. Chapter ix applications of the method of multipliers to variational inequalities. *Studies in mathematics and its applications*, 15:299–331, 1983.

[21] W. L. Hare and A. S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

[22] M. Kadkhodaie, K. Christakopoulou, M. Sanjabi, and A. Banerjee. Accelerated alternating direction method of multipliers. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 497–506. ACM, 2015.

[23] J. M. Lee. *Smooth manifolds*. Springer, 2003.

[24] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2003.

[25] J. Liang. *Convergence rates of first-order operator splitting methods*. PhD thesis, Normandie Université; GREYC CNRS UMR 6072, 2016.

[26] J. Liang, J. Fadili, and G. Peyré. Local linear convergence of Forward–Backward under partial smoothness. In *Advances in Neural Information Processing Systems*, pages 1970–1978, 2014.

[27] J. Liang, J. Fadili, and G. Peyré. Activity identification and local linear convergence of Forward–Backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017.

[28] J. Liang, J. Fadili, and G. Peyré. Local convergence properties of Douglas–Rachford and alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 172(3):874–913, 2017.

[29] M. Mešina. Convergence acceleration for the iterative solution of the equations x= ax+ f. *Computer Methods in Applied Mechanics and Engineering*, 10(2):165–173, 1977.

[30] S. A. Miller and J. Malick. Newton methods for nonsmooth convex minimization: connections among-Lagrangian, Riemannian Newton and SQP methods. *Mathematical programming*, 104(2-3):609–633, 2005.

[31] Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[32] I. Pejcic and C. N. Jones. Accelerated admm based on accelerated douglas-rachford splitting. In *2016 European Control Conference (ECC)*, pages 1952–1957. Ieee, 2016.

[33] L. F. Richardson and J. A. Gaunt. Viii. the deferred approach to the limit. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 226(636-646):299–361, 1927.

[34] D. Scieur, A. d'Aspremont, and F. Bach. Regularized nonlinear acceleration. In *Advances In Neural Information Processing Systems*, pages 712–720, 2016.

[35] D. Shanks. Non-linear transformations of divergent and slowly convergent sequences. *Journal of Mathematics and Physics*, 34(1-4):1–42, 1955.

[36] A. Sidi. *Practical extrapolation methods: Theory and applications*, volume 10. Cambridge University Press, 2003.

[37] A. Sidi. *Vector extrapolation methods with applications*, volume 17. SIAM, 2017.

[38] A. Sidi, W. F. Ford, and D. A. Smith. Acceleration of convergence of vector sequences. *SIAM Journal on Numerical Analysis*, 23(1):178–196, 1986.

[39] A. Sidi and Y. Shapira. Upper bounds for convergence rates of acceleration methods with initial iterations. *Numerical Algorithms*, 18(2):113–132, 1998.

[40] S. Vaiter, G. Peyré, and J. Fadili. Model consistency of partly smooth regularizers. *IEEE Transactions on Information Theory*, 64(3):1725–1737, 2018.

[41] P. Wynn. Acceleration techniques for iterated vector and matrix problems. *Mathematics of Computation*, 16(79):301–322, 1962.

Appendix

The organisation of the appendix is as follows: the proofs of the main results of the paper are contained in Sections A-C, where in Section A some preliminary result on angles between subspaces and Riemannian geometry are provide, in Section B the proofs for the trajectory of ADMM are provided, and lastly in in Section C we provide proofs on A^{3}DMM.

A Preliminaries

A.1 Polynomial extrapolation

Minimal polynomial extrapolation (MPE) [13]: Given $\{z_{k-j}\}_{j=0}^{q+1}$, let $\{v_{k-j}\}_{j=0}^{q}$ be the difference vectors, where $v_j \overset{\text{def}}{=} z_j - z_{j-1}$. Define $V_k = \begin{bmatrix} v_k & \cdots & v_{k-q} \end{bmatrix}$.

1. Let $\{c_j\}_{j=1}^{q} \in \operatorname{argmin}_{c\in\mathbb{R}^q} \|V_{k-1}c - v_k\|$, define $c_0 \overset{\text{def}}{=} 1$ and $\gamma_i = c_i / \sum_{i=0}^{q} c_i$ for $i = 0, \ldots, q$.
2. The extrapolated point is then defined to be $\bar{z}_k \overset{\text{def}}{=} \sum_{i=0}^{q} \gamma_i z_{k-i-1}$.

Reduced rank extrapolation (RRE) [18, 29] is obtained by replacing the first step by

$$\{\gamma_j\}_{j=0}^{q} \in \operatorname{argmin}_{\gamma\in\mathbb{R}^{q+1}} \|V_k\gamma\| \text{ subject to } \sum\nolimits_i \gamma_i = 1.$$

The motivation for the use of such methods for the acceleration of fixed point sequences $x_{k+1} = \mathcal{F}(z_k)$ come from considering the spectral properties of the linearisation around the limit point. In particular, if $z^\star$ is the limit point and $z_{k+1} - z^\star = T(z_k - z^\star)$ where $T \in \mathbb{R}^{d\times d}$ and q is the order of the minimal polynomial of T with respect to $z_{k-q-1} - z^\star$ (i.e. q is the monic polynomial of least degree such that $P(T)(z_{k-q-1} - z^\star) = 0$), then one can show that $\bar{z}_k = z^\star$. We refer to [38, 39, 37] for details on these methods and their acceleratioon guarantees.

A.2 Angle between subspaces

Let T_1, T_2 be two subspaces, and without the loss of generality, assume

$$1 \le p \overset{\text{def}}{=} \dim(T_1) \le q \overset{\text{def}}{=} \dim(T_2) \le n - 1.$$

Definition A.1 (Principal angles). The principal angles $\theta_k \in [0, \frac{\pi}{2}]$, $k = 1, \ldots, p$ between subspaces T_1 and T_2 are defined by, with $u_0 = v_0 \overset{\text{def}}{=} 0$, and

$$\cos(\theta_k) \overset{\text{def}}{=} \langle u_k, v_k\rangle = \max\langle u, v\rangle \text{ s.t. } u \in T_1, v \in T_2, \|u\| = 1, \|v\| = 1, \\ \langle u, u_i\rangle = \langle v, v_i\rangle = 0,\ i = 0, \cdots, k-1.$$

The principal angles θ_k are unique and satisfy $0 \le \theta_1 \le \theta_2 \le \cdots \le \theta_p \le \pi/2$.

Definition A.2 (Friedrichs angle). The Friedrichs angle $\theta_F \in]0, \frac{\pi}{2}]$ between T_1 and T_2 is

$$\cos\left(\theta_F(T_1, T_2)\right) \overset{\text{def}}{=} \max\langle u, v\rangle \text{ s.t. } u \in T_1 \cap (T_1 \cap T_2)^\perp, \|u\| = 1,\ v \in T_2 \cap (T_1 \cap T_2)^\perp, \|v\| = 1.$$

The following lemma shows the relation between the Friedrichs and principal angles, whose proof can be found in [6, Proposition 3.3].

Lemma A.3 (Principal angles and Friedrichs angle). *The Friedrichs angle is exactly θ_{d+1} where $d \overset{\text{def}}{=} \dim(T_1 \cap T_2)$. Moreover, $\theta_F(T_1, T_2) > 0$.*

A.3 Riemannian Geometry

Let $\mathcal{M}$ be a C^2-smooth embedded submanifold of $\mathbb{R}^n$ around a point x. With some abuse of terminology, we shall state C^2-manifold instead of C^2-smooth embedded submanifold of $\mathbb{R}^n$. The natural embedding of a submanifold $\mathcal{M}$ into $\mathbb{R}^n$ permits to define a Riemannian structure and to introduce geodesics on $\mathcal{M}$, and we simply say $\mathcal{M}$ is a Riemannian manifold. We denote respectively $\mathcal{T}_{\mathcal{M}}(x)$ and $\mathcal{N}_{\mathcal{M}}(x)$ the tangent and normal space of $\mathcal{M}$ at point near x in $\mathcal{M}$.

Exponential map Geodesics generalize the concept of straight lines in $\mathbb{R}^n$, preserving the zero acceleration characteristic, to manifolds. Roughly speaking, a geodesic is locally the shortest path between two points on $\mathcal{M}$. We denote by $\mathfrak{g}(t; x, h)$ the value at $t \in \mathbb{R}$ of the geodesic starting at $\mathfrak{g}(0; x, h) = x \in \mathcal{M}$ with velocity $\dot{\mathfrak{g}}(t; x, h) = \frac{d\mathfrak{g}}{dt}(t; x, h) = h \in \mathcal{T}_{\mathcal{M}}(x)$ (which is uniquely defined). For every $h \in \mathcal{T}_{\mathcal{M}}(x)$, there exists an interval I around 0 and a unique geodesic $\mathfrak{g}(t; x, h) : I \to \mathcal{M}$ such that $\mathfrak{g}(0; x, h) = x$ and $\dot{\mathfrak{g}}(0; x, h) = h$. The mapping

$$\mathrm{Exp}_x : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{M}, \;\; h \mapsto \mathrm{Exp}_x(h) = \mathfrak{g}(1; x, h),$$

is called *Exponential map*. Given $x, x' \in \mathcal{M}$, the direction $h \in \mathcal{T}_{\mathcal{M}}(x)$ we are interested in is such that

$$\mathrm{Exp}_x(h) = x' = \mathfrak{g}(1; x, h).$$

Parallel translation Given two points $x, x' \in \mathcal{M}$, let $\mathcal{T}_{\mathcal{M}}(x), \mathcal{T}_{\mathcal{M}}(x')$ be their corresponding tangent spaces. Define

$$\tau : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x'),$$

the parallel translation along the unique geodesic joining x to x', which is isomorphism and isometry w.r.t. the Riemannian metric.

Riemannian gradient and Hessian For a vector $v \in \mathcal{N}_{\mathcal{M}}(x)$, the Weingarten map of $\mathcal{M}$ at x is the operator $\mathfrak{W}_x(\cdot, v) : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x)$ defined by

$$\mathfrak{W}_x(\cdot, v) = -\mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x)} \mathrm{d}V[h],$$

where V is any local extension of v to a normal vector field on $\mathcal{M}$. The definition is independent of the choice of the extension V, and $\mathfrak{W}_x(\cdot, v)$ is a symmetric linear operator which is closely tied to the second fundamental form of $\mathcal{M}$, see [15, Proposition II.2.1].

Let G be a real-valued function which is C^2 along the $\mathcal{M}$ around x. The covariant gradient of G at $x' \in \mathcal{M}$ is the vector $\nabla_{\mathcal{M}} G(x') \in \mathcal{T}_{\mathcal{M}}(x')$ defined by

$$\langle \nabla_{\mathcal{M}} G(x'),\, h \rangle = \frac{d}{dt} G\big(\mathcal{P}_{\mathcal{M}}(x' + th)\big)\big|_{t=0}, \;\; \forall h \in \mathcal{T}_{\mathcal{M}}(x'),$$

where $\mathcal{P}_{\mathcal{M}}$ is the projection operator onto $\mathcal{M}$. The covariant Hessian of G at x' is the symmetric linear mapping $\nabla^2_{\mathcal{M}} G(x')$ from $\mathcal{T}_{\mathcal{M}}(x')$ to itself which is defined as

$$\langle \nabla^2_{\mathcal{M}} G(x')h,\, h \rangle = \frac{d^2}{dt^2} G\big(\mathcal{P}_{\mathcal{M}}(x' + th)\big)\big|_{t=0}, \;\; \forall h \in \mathcal{T}_{\mathcal{M}}(x'). \tag{A.1}$$

This definition agrees with the usual definition using geodesics or connections [30]. Now assume that $\mathcal{M}$ is a Riemannian embedded submanifold of $\mathbb{R}^n$, and that a function G has a C^2-smooth restriction on $\mathcal{M}$. This can be characterized by the existence of a C^2-smooth extension (representative) of G, *i.e.* a C^2-smooth function $\widetilde{G}$ on $\mathbb{R}^n$ such that $\widetilde{G}$ agrees with G on $\mathcal{M}$. Thus, the Riemannian gradient $\nabla_{\mathcal{M}} G(x')$ is also given by

$$\nabla_{\mathcal{M}} G(x') = \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla \widetilde{G}(x'), \tag{A.2}$$

and $\forall h \in \mathcal{T}_{\mathcal{M}}(x')$, the Riemannian Hessian reads

$$\begin{aligned} \nabla^2_{\mathcal{M}} G(x')h &= \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x')} \mathrm{d}(\nabla_{\mathcal{M}} G)(x')[h] = \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x')} \mathrm{d}\big(x' \mapsto \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla_{\mathcal{M}} \widetilde{G}\big)[h] \\ &= \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla^2 \widetilde{G}(x')h + \mathfrak{W}_{x'}\big(h, \mathcal{P}_{\mathcal{N}_{\mathcal{M}}(x')} \nabla \widetilde{G}(x')\big), \end{aligned} \tag{A.3}$$

where the last equality comes from [1, Theorem 1]. When $\mathcal{M}$ is an affine or linear subspace of $\mathbb{R}^n$, then obviously $\mathcal{M} = x + \mathcal{T}_{\mathcal{M}}(x)$, and $\mathfrak{W}_{x'}(h, \mathcal{P}_{\mathcal{N}_{\mathcal{M}}(x')} \nabla \widetilde{G}(x')) = 0$, hence (A.3) reduces to

$$\nabla^2_{\mathcal{M}} G(x') = \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x')} \nabla^2 \widetilde{G}(x') \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x')}.$$

See [23, 15] for more materials on differential and Riemannian manifolds.

A.4 Preparatory lemmas

The following lemmas characterise the parallel translation and the Riemannian Hessian of nearby points in $\mathcal{M}$.

Lemma A.4 ([26, Lemma 5.1]). *Let $\mathcal{M}$ be a C^2-smooth manifold around x. Then for any $x' \in \mathcal{M} \cap \mathcal{N}$, where $\mathcal{N}$ is a neighbourhood of x, the projection operator $\mathcal{P}_{\mathcal{M}}(x')$ is uniquely valued and C^1 around x, and thus*

$$x' - x = \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x)}(x' - x) + o(\|x' - x\|).$$

If moreover $\mathcal{M} = x + \mathcal{T}_{\mathcal{M}}(x)$ is an affine subspace, then $x' - x = \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x)}(x' - x)$.

Lemma A.5 ([27, Lemma B.1]). *Let $x \in \mathcal{M}$, and x_k a sequence converging to x in $\mathcal{M}$. Denote $\tau_k : \mathcal{T}_{\mathcal{M}}(x_k) \to \mathcal{T}_{\mathcal{M}}(x)$ be the parallel translation along the unique geodesic joining x to x_k. Then, for any bounded vector $u \in \mathbb{R}^n$, we have*

$$(\tau_k \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x_k)} - \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x)})u = o(\|u\|).$$

The Riemannian gradient and Hessian of partly smooth functions are covered by the lemma below.

Lemma A.6 ([27, Lemma B.2]). *Let x, x' be two close points in $\mathcal{M}$, denote $\tau : \mathcal{T}_{\mathcal{M}}(x') \to \mathcal{T}_{\mathcal{M}}(x)$ the parallel translation along the unique geodesic joining x to x'. The Riemannian Taylor expansion of $R \in C^2(\mathcal{M})$ around x reads,*

$$\tau \nabla_{\mathcal{M}} R(x') = \nabla_{\mathcal{M}} R(x) + \nabla^2_{\mathcal{M}} R(x) \mathcal{P}_{\mathcal{T}_{\mathcal{M}}(x)}(x' - x) + o(\|x' - x\|).$$

Lemma A.7 (Riemannian gradient and Hessian). *If $R \in \mathrm{PSF}_x(\mathcal{M}_x)$, then for any point $x' \in \mathcal{M}_x$ near x*

$$\nabla_{\mathcal{M}_x} R(x') = \mathcal{P}_{T_{x'}}(\partial R(x')),$$

and this does not depend on the smooth representation of R on $\mathcal{M}_x$. In turn, for all $h \in T_{x'}$, let $\widetilde{R}$ be a smooth representative of R on $\mathcal{M}_x$,

$$\nabla^2_{\mathcal{M}_x} R(x')h = \mathcal{P}_{T_{x'}} \nabla^2 \widetilde{R}(x')h + \mathfrak{W}_{x'}\big(h, \mathcal{P}_{T^{\perp}_{x'}} \nabla \widetilde{R}(x')\big),$$

where $\mathfrak{W}_x(\cdot, \cdot) : T_x \times T_x^{\perp} \to T_x$ is the Weingarten map of $\mathcal{M}_x$ at x.

A.5 Linearisation of proximal mapping

In this part, we present one fundamental result led by partial smoothness, the linearisation of proximal mapping. We first discuss the property of the Riemannian Hessian of a partly smooth function. Let $R \in \Gamma_0(\mathbb{R}^n)$ be partly smooth at $\bar{x}$ relative to $\mathcal{M}_{\bar{x}}$ and $\bar{u} \in \partial R(\bar{x})$, define the following smooth perturbation of R

$$\overline{R}(x) \stackrel{\text{def}}{=} R(x) - \langle x, \bar{u} \rangle,$$

whose Riemannian Hessian at $\bar{x}$ reads $H_{\overline{R}} \stackrel{\text{def}}{=} \mathcal{P}_{T_{\bar{x}}} \nabla^2_{\mathcal{M}_{\bar{x}}} \overline{R}(\bar{x}) \mathcal{P}_{T_{\bar{x}}}$.

Lemma A.8 ([27, Lemma 4.2]). *Let $R \in \Gamma_0(\mathbb{R}^n)$ be partly smooth at $\bar{x}$ relative to $\mathcal{M}_{\bar{x}}$, then $H_{\overline{R}}$ is symmetric positive semi-definite if either of the following is true:*

- *$\bar{u} \in \mathrm{ri}(\partial R(\bar{x}))$ is non-denegerate.*
- *$\mathcal{M}_{\bar{x}}$ is an affine subspace.*

In turn, $\mathrm{Id} + H_{\overline{R}}$ is invertible and $M_{\overline{R}} \stackrel{\text{def}}{=} (\mathrm{Id} + H_{\overline{R}})^{-1}$ is symmetric positive definite with all eigenvalues in $]0, 1]$.

One consequence of Lemma A.8 is that, we can linearise the generalised proximal mapping. For the sake of generality, let $\gamma > 0$, $R \in \Gamma_0(\mathbb{R}^n)$ and $A \in \mathbb{R}^{p \times n}$, define the following generalised proximal mapping

$$\mathrm{prox}^A_{\gamma R}(\cdot) \stackrel{\text{def}}{=} \mathrm{argmin}_{x \in \mathbb{R}^n} \gamma R(x) + \frac{1}{2}\|Ax - \cdot\|^2.$$

Clearly, $\mathrm{prox}^A_{\gamma R}$ is a single-valued mapping when A has full column rank. Define $A_{T_{\bar{x}}} = A \circ \mathcal{P}_{T_{\bar{x}}}$, which has full column rank owing to A. Hence $A^T_{T_{\bar{x}}} A_{T_{\bar{x}}}$ is invertible. Denote

$$M_{\overline{R}} = A_{T_{\bar{x}}}(\mathrm{Id} + (A^T_{T_{\bar{x}}} A_{T_{\bar{x}}})^{-1} H_{\overline{R}})^{-1} (A^T_{T_{\bar{x}}} A_{T_{\bar{x}}})^{-1} A^T_{T_{\bar{x}}}.$$

Lemma A.9. *Let function $R \in \Gamma_0(\mathbb{R}^n)$ be partly smooth at the point $\bar{x}$ relative to the manifold $\mathcal{M}_{\bar{x}}$ and $\bar{u} \in \mathrm{ri}(\partial R(\bar{x}))$. Suppose that there exists $\gamma > 0$, full column rank $A \in \mathbb{R}^{p\times n}$ and $\bar{w} \in \mathbb{R}^p$ such that $\bar{x} = \mathrm{prox}^A_{\gamma R}(\bar{w})$ and $\bar{u} = -A^T(A\bar{x} - \bar{w})/\gamma$. Let $\{w_k\}_{k\in\mathbb{N}}$ be a sequence such that $w_k \to \bar{w}$ and $x_k = \mathrm{prox}^A_{\gamma R}(w_k) \to \bar{x}$, then for all k large enough, there holds*

$$A_{T_{\bar{x}}}(x_k - x_{k-1}) = M_{\overline{R}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|). \tag{A.4}$$

Remark A.10. When $A = \mathrm{Id}$, then $\mathrm{prox}^A_{\gamma R}$ reduces to the standard proximal mapping, and (A.4) simplifies to

$$x_k - x_{k-1} = \mathcal{P}_{T_{\bar{x}}}\big(\mathrm{Id} + H_{\overline{R}}\big)^{-1}\mathcal{P}_{T_{\bar{x}}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|).$$

In [25] and references therein, to study the local linear convergence of first-order methods, linearisation with respect to the limiting points is provided, that is

$$x_k - \bar{x} = \mathcal{P}_{T_{\bar{x}}}\big(\mathrm{Id} + H_{\overline{R}}\big)^{-1}\mathcal{P}_{T_{\bar{x}}}(w_k - \bar{w}) + o(\|w_k - \bar{w}\|).$$

Proof. Since R is proper convex and lower semi-continuous, we have $R(x_k) \to R(\bar{x})$ and $\partial R(x_k) \ni u_k = -A^T(Ax_k - w_k)/\gamma \to \bar{u} \in \mathrm{ri}(\partial R(\bar{x}))$, and we have $x_k \in \mathcal{M}_{\bar{x}}$ owing to [21, Theorem 5.3] and $u_k \in \mathrm{ri}(\partial R(x_k))$ owing to [40] for all k large enough.

Denote $T_{x_k}, T_{x_{k-1}}$ the tangent spaces of $\mathcal{M}_{\bar{x}}$ at x_k and x_{k-1}. Denote $\tau_k : T_{x_k} \to T_{x_{k-1}}$ the parallel translation along the unique geodesic on $\mathcal{M}_{\bar{x}}$ joining x_k to x_{k-1}.

From the definition of x_k, we get

$$u_k \overset{\text{def}}{=} -A^T(Ax_k - w_k) \in \gamma\partial R(x_k) \quad \text{and} \quad u_{k-1} \overset{\text{def}}{=} -A^T(Ax_{k-1} - w_{k-1}) \in \gamma\partial R(x_{k-1}).$$

Projecting on the corresponding tangent spaces, applying Lemma A.7 and the parallel translation τ_k leads to

$$\begin{aligned}
\gamma\tau_k\nabla_{\mathcal{M}_{\bar{x}}}R(x_k) &= \tau_k\mathcal{P}_{T^R_{x_k}}(u_k) = \mathcal{P}_{T^R_{x_{k-1}}}(u_k) + \big(\tau_k\mathcal{P}_{T^J_{x_k}} - \mathcal{P}_{T^R_{x_{k-1}}}\big)(u_k),\\
\gamma\nabla_{\mathcal{M}_{\bar{x}}}R(x_{k-1}) &= \mathcal{P}_{T^R_{x_{k-1}}}(u_{k-1}).
\end{aligned}$$

The difference of the above two equalities leads to

$$\begin{aligned}
&\gamma\tau_k\nabla_{\mathcal{M}_{\bar{x}}}R(x_k) - \gamma\nabla_{\mathcal{M}_{\bar{x}}}R(x_{k-1}) - \big(\tau_k\mathcal{P}_{T^J_{x_k}} - \mathcal{P}_{T^R_{x_{k-1}}}\big)(u_{k-1})\\
&= \mathcal{P}_{T^R_{x_{k-1}}}(u_k - u_{k-1}) + \big(\tau_k\mathcal{P}_{T^J_{x_k}} - \mathcal{P}_{T^R_{x_{k-1}}}\big)(u_k - u_{k-1}).
\end{aligned} \tag{A.5}$$

From the monotonicity from subdifferential, *i.e.* $\langle u_k - u_{k-1},\, x_k - x_{k-1}\rangle \geq 0$, we get

$$\langle A^TA(x_k - x_{k-1}),\, x_k - x_{k-1}\rangle \leq \langle A^T(w_k - w_{k-1}),\, x_k - x_{k-1}\rangle \leq \|A\|\|w_k - w_{k-1}\|\|x_k - x_{k-1}\|.$$

Since A has full column rank, then A^TA is symmetric positive definite, and there exists $\kappa > 0$ such that $\kappa\|x_k - x_{k-1}\|^2 \leq \langle A^TA(x_k - x_{k-1}),\, x_k - x_{k-1}\rangle$. Back to the above inequality, we get $\|x_k - x_{k-1}\| \leq \frac{\|A\|}{\kappa}\|w_k - w_{k-1}\|$. Therefore for $\|u_k - u_{k-1}\|$, we get

$$\begin{aligned}
\|u_k - u_{k-1}\| = \|A^T(Ax_k - w_k) - A^T(Ax_{k-1} - w_{k-1})\| &\leq \|A\|^2\|x_k - x_{k-1}\| + \|A\|\|w_k - w_{k-1}\|\\
&\leq \big(\tfrac{\|A\|^3}{\kappa} + \|A\|\big)\|w_k - w_{k-1}\|.
\end{aligned}$$

As a result, owing to Lemma A.5, we have for the term $\big(\tau_k\mathcal{P}_{T^J_{x_k}} - \mathcal{P}_{T^R_{x_{k-1}}}\big)(u_k - u_{k-1})$ in (A.5)

$$\big(\tau_k\mathcal{P}_{T^J_{x_k}} - \mathcal{P}_{T^R_{x_{k-1}}}\big)(u_k - u_{k-1}) = o(\|w_k - w_{k-1}\|).$$

Define $\overline{R}_{k-1}(x) \overset{\text{def}}{=} \gamma R(x) - \langle x,\, u_{k-1}\rangle$ and $H_{\overline{R},k-1} \overset{\text{def}}{=} \mathcal{P}_{T_{x_{k-1}}}\nabla^2_{\mathcal{M}_{\bar{x}}}\overline{R}(x_{k-1})\mathcal{P}_{T_{x_{k-1}}}$, then with Lemma A.6 the Riemannian Taylor expansion, we have for the first line of (A.5)

$$\begin{aligned}
&\gamma\tau_k\nabla_{\mathcal{M}_{\bar{x}}}R(x_k) - \gamma\nabla_{\mathcal{M}_{\bar{x}}}R(x_{k-1}) - \big(\tau_k\mathcal{P}_{T_{x_k}} - \mathcal{P}_{T_{x_{k-1}}}\big)(u_{k-1})\\
&= \tau_k\big(\gamma\nabla_{\mathcal{M}_{\bar{x}}}R(x_k) - \mathcal{P}_{T^J_{x_k}}(u_{k-1})\big) - \big(\gamma\nabla_{\mathcal{M}_{\bar{x}}}R(x_{k-1}) - \mathcal{P}_{T_{x_{k-1}}}(u_{k-1})\big)\\
&= \tau_k\nabla_{\mathcal{M}_{\bar{x}}}\overline{R}_{k-1}(x_k) - \nabla_{\mathcal{M}_{\bar{x}}}\overline{R}_{k-1}(x_{k-1})\\
&= H_{\overline{R},k-1}(x_k - x_{k-1}) + o(\|x_k - x_{k-1}\|)\\
&= H_{\overline{R},k-1}(x_k - x_{k-1}) + o(\|w_k - w_{k-1}\|).
\end{aligned} \tag{A.6}$$

Back to (A.5), we get

$$H_{\overline{R},k-1}(x_k - x_{k-1}) = \mathcal{P}_{T^R_{x_{k-1}}}(u_k - u_{k-1}) + o(\|w_k - w_{k-1}\|). \tag{A.7}$$

Define $\overline{R}(x) \overset{\text{def}}{=} \gamma R(x) - \langle x,\, \bar{u}\rangle$ and $H_{\overline{R}} = \mathcal{P}_{T_{\bar{x}}} \nabla^2_{\mathcal{M}_{\bar{x}}} \overline{R}(\bar{x}) \mathcal{P}_{T_{\bar{x}}}$, then from (A.7) that

$$\begin{aligned} &H_{\overline{R}}(x_k - x_{k-1}) + \big(H_{\overline{R},k-1} - H_{\overline{R}}\big)(x_k - x_{k-1}) \\ &= \mathcal{P}_{T_{\bar{x}}}(u_k - u_{k-1}) + \big(H_{\overline{R},k-1} - H_{\overline{R}}\big)(u_k - u_{k-1}) + o(\|w_k - w_{k-1}\|). \end{aligned} \tag{A.8}$$

Owing to continuity, we have $H_{\overline{R},k-1} \to H_{\overline{R}}$ and $\mathcal{P}_{T_{x_{k-1}}} \to \mathcal{P}_{T_{\bar{x}}}$, and

$$\begin{aligned} \lim_{k\to+\infty} \frac{\|(H_{\overline{R},k-1} - H_{\overline{R}})(x_k - x_{k-1})\|}{\|x_k - x_{k-1}\|} &\le \lim_{k\to+\infty} \frac{\|H_{\overline{R},k-1} - H_{\overline{R}}\|\|x_k - x_{k-1}\|}{\|x_k - x_{k-1}\|} = \lim_{k\to+\infty} \|H_{\overline{R},k-1} - H_{\overline{R}}\| = 0, \\ \lim_{k\to+\infty} \frac{\|(\mathcal{P}_{T_{x_{k-1}}} - \mathcal{P}_{T_{\bar{x}}})(w_k - w_{k-1})\|}{\|w_k - w_{k-1}\|} &\le \lim_{k\to+\infty} \frac{\|\mathcal{P}_{T_{x_{k-1}}} - \mathcal{P}_{T_{\bar{x}}}\|\|w_k - w_{k-1}\|}{\|w_k - w_{k-1}\|} = \lim_{k\to+\infty} \|\mathcal{P}_{T_{x_{k-1}}} - \mathcal{P}_{T_{\bar{x}}}\| = 0, \\ \lim_{k\to+\infty} \frac{\|(\mathcal{P}_{T_{x_{k-1}}} - \mathcal{P}_{T_{\bar{x}}})(x_k - x_{k-1})\|}{\|x_k - x_{k-1}\|} &= 0. \end{aligned}$$

Combining this with the definition of u_k, the fact that $x_k - x_{k-1} = \mathcal{P}_{T_{\bar{x}}}(x_k - x_{k-1}) + o(\|x_k - x_{k-1}\|)$ from Lemma A.4, and denoting $A_{T_{\bar{x}}} = A \circ \mathcal{P}_{T_{\bar{x}}}$, equation (A.8) can be written as

$$\begin{aligned} H_{\overline{R}}(x_k - x_{k-1}) &= \mathcal{P}_{T_{\bar{x}}}(u_k - u_{k-1}) + o(\|w_k - w_{k-1}\|) \\ &= -\mathcal{P}_{T_{\bar{x}}}(A^T(Ax_k - w_k) - A^T(Ax_{k-1} - w_{k-1})) + o(\|w_k - w_{k-1}\|) \\ &= -\mathcal{P}_{T_{\bar{x}}} A^T A(x_k - x_{k-1}) + \mathcal{P}_{T_{\bar{x}}} A^T(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|) \\ &= -A^T_{T_{\bar{x}}} A_{T_{\bar{x}}}(x_k - x_{k-1}) + A^T_{T_{\bar{x}}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|) \end{aligned} \tag{A.9}$$

Since A has full rank, so is $A_{T_{\bar{x}}}$. Hence $A^T_{T_{\bar{x}}} A_{T_{\bar{x}}}$ is invertible and from above we have

$$\big(\mathrm{Id} + (A^T_{T_{\bar{x}}} A_{T_{\bar{x}}})^{-1} H_{\overline{R}}\big)(x_k - x_{k-1}) = (A^T_{T_{\bar{x}}} A_{T_{\bar{x}}})^{-1} A^T_{T_{\bar{x}}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|).$$

Denote $M_{\overline{R}} = A_{T_{\bar{x}}}(\mathrm{Id} + (A^T_{T_{\bar{x}}} A_{T_{\bar{x}}})^{-1} H_{\overline{R}})^{-1} (A^T_{T_{\bar{x}}} A_{T_{\bar{x}}})^{-1} A^T_{T_{\bar{x}}}$, then

$$A_{T_{\bar{x}}}(x_k - x_{k-1}) = M_{\overline{R}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|), \tag{A.10}$$

which concludes the proof. □

B Trajectory of ADMM

In this section, we first provide the fixed-point characterisation of ADMM based on the equivalence between ADMM and Douglas–Rachford, and then present the proofs for the trajectory of ADMM.

B.1 Fixed-point characterisation and convergence of ADMM

It is well-known that ADMM is equivalent to applying Douglas–Rachford splitting [17] to solve the dual problem of ($\mathcal{P}$) which reads

$$\max_{\psi\in\mathbb{R}^p} -\big(R^*(-A^T\psi) + J^*(-B^T\psi) + \langle \psi,\, b\rangle\big), \tag{$\mathcal{D}_{\text{ADMM}}$}$$

where $R^*(v) \overset{\text{def}}{=} \sup_{x\in\mathbb{R}^n} (\langle x,\, v\rangle - R(x))$ is called the Fenchel conjugate, or simply conjugate, of R. Below we first recall the equivalence between ADMM and Douglas–Rachford which was first established in [20], and then use the convergence of Douglas–Rachford splitting method which is well established in the literature [5] to conclude the convergence of ADMM.

- For the update of x_k, denote $u_k = \gamma(Ax_k + By_{k-1} - b) + \psi_{k-1}$ and $z_k = \psi_k - \gamma By_k + \gamma b$. Since A has full column rank, we have x_k is the unique minimiser of $R(x) + \frac{\gamma}{2}\|Ax + By_{k-1} - b + \frac{1}{\gamma}\psi_{k-1}\|^2$. Let R^* be the conjugate of R, then owing to duality, we get

$$\begin{aligned}
& x_k = \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax + By_{k-1} - b + \tfrac{1}{\gamma}\psi_{k-1}\|^2 \\
\iff\ & 0 \in \partial R(x_k) + \gamma A^T\big(Ax_k + By_{k-1} - b + \tfrac{1}{\gamma}\psi_{k-1}\big) \\
\iff\ & -A^T u_k \in \partial R(x_k) \\
\iff\ & x_k \in \partial R^*(-A^T u_k) \\
\iff\ & u_k - \gamma Ax_k \in u_k + \gamma\partial(R^* \circ -A^T)(u_k) \\
\iff\ & u_k = \big(\mathrm{Id} + \gamma\partial(R^* \circ -A^T)\big)^{-1}(u_k - \gamma Ax_k) \\
\iff\ & u_k = \big(\mathrm{Id} + \gamma\partial(R^* \circ -A^T)\big)^{-1}(2\psi_{k-1} - z_{k-1}).
\end{aligned}$$

- For the update of y_k, the full column rank of B also ensures that y_k is the unique minimiser of $J(y) + \frac{\gamma}{2}\|Ax_k + By - b + \frac{1}{\gamma}\psi_{k-1}\|^2$. Since $\psi_k = \psi_{k-1} + \gamma(Ax_k + By_k - b)$, then

$$\begin{aligned}
& y_{k+1} = \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|Ax_{k+1} + By - b + \tfrac{1}{\gamma}\psi_{k-1}\|^2 \\
\iff\ & 0 \in \partial J(y_k) + \gamma B^T\big(Ax_k + By_k - b + \tfrac{1}{\gamma}\psi_{k-1}\big) \\
\iff\ & -B^T\psi_k \in \partial J(y_k) \\
\iff\ & y_k \in \partial J^*(-B^T\psi_k) \\
\iff\ & \psi_k - \gamma By_k \in \psi_k + \gamma\partial(J^* \circ -B^T)(\psi_k) \\
\iff\ & \psi_k = \big(\mathrm{Id} + \gamma\partial(J^* \circ -B^T)\big)^{-1}(\psi_k - \gamma By_k) \\
\iff\ & \psi_k = \big(\mathrm{Id} + \gamma\partial(J^* \circ -B^T)\big)^{-1}(z_k - \gamma b).
\end{aligned}$$

- Summing up the above two relations we get

$$\begin{aligned}
u_k &= \big(\mathrm{Id} + \gamma\partial(R^* \circ -A^T)\big)^{-1}(2\psi_{k-1} - z_{k-1}), \\
z_k &= z_{k-1} + u_k - \psi_{k-1}, \\
\psi_k &= \big(\mathrm{Id} + \gamma\partial(J^* \circ -B^T)\big)^{-1}(z_k - \gamma b),
\end{aligned} \tag{B.1}$$

  which is exactly the iteration of Douglas–Rachford splitting algorithm when applied to solving the dual problem ($\mathcal{D}_{\mathrm{ADMM}}$).

Define the following operator

$$\mathcal{F} = \tfrac{1}{2}\mathrm{Id} + \tfrac{1}{2}\Big(2\big(\mathrm{Id} + \gamma\partial(R^* \circ -A^T)\big)^{-1} - \mathrm{Id}\Big)\Big(2\big(\mathrm{Id} + \gamma\partial(J^* \circ -B^T)\big)^{-1} - \mathrm{Id}\Big),$$

then (B.1) can be written as the fixed-point iteration in terms of z_k, that is

$$z_k = \mathcal{F}(z_{k-1}).$$

It should be noted that for z_k we have $z_k = \psi_k - \gamma By_k + \gamma b = \psi_{k-1} + \gamma Ax_k$ which is the same as in (1.2). Owing to [5], we have that $\mathcal{F}$ is firmly non-expansive with the set of fixed-points $\mathrm{fix}(\mathcal{F})$ being non-empty, and there exists a fixed-point $z^\star \in \mathrm{fix}(\mathcal{F})$ such that $z_k \to z^\star$ which concludes the convergence of $\{z_k\}_{k\in\mathbb{N}}$. Then we have u_k, ψ_k converging to $\psi^\star = \big(\mathrm{Id} + \gamma\partial(J^* \circ -B^T)\big)^{-1}(z^\star - \gamma b)$ which is a dual solution of the problem ($\mathcal{D}_{\mathrm{ADMM}}$). The convergence of the primal ADMM sequences $\{x_k\}_{k\in\mathbb{N}}$ and $\{y_k\}_{k\in\mathbb{N}}$ follows immediately.

Owing to the above equivalence between ADMM and Douglas–Rachford splitting, we get the following relations

$$\begin{aligned}
\|z_k - z_{k-1}\| &\le \|z_{k-1} - z_{k-2}\|, \\
\|\psi_k - \psi_{k-1}\| &\le \|z_k - z_{k-1}\| \le \|z_{k-1} - z_{k-2}\|, \\
\|u_k - u_{k-1}\| &\le \|2\psi_{k-1} - z_{k-1} - 2\psi_{k-2} + z_{k-2}\| \le 3\|z_{k-1} - z_{k-2}\|, \\
\gamma\|Ax_k - Ax_{k-1}\| &\le \|z_k - z_{k-1}\| + \|\psi_{k-1} - \psi_{k-2}\| \le 2\|z_{k-1} - z_{k-2}\|, \\
\gamma\|By_k - By_{k-1}\| &\le \|z_k - z_{k-1}\| + \|\psi_k - \psi_{k-1}\| \le 2\|z_{k-1} - z_{k-2}\|,
\end{aligned} \tag{B.2}$$

which are needed in the proofs below.

B.2 Trajectory of ADMM: both R, J are non-smooth

Given a saddle point $(x^\star, y^\star, \psi^\star)$ of $\mathcal{L}(x, y; \psi)$, the first-order optimality condition entails $-A^T\psi^\star \in \partial R(x^\star)$ and $-B^T\psi^\star \in \partial J(y^\star)$. Below we impose a stronger condition

$$-A^T\psi^\star \in \mathrm{ri}\big(\partial R(x^\star)\big) \quad \text{and} \quad -B^T\psi^\star \in \mathrm{ri}\big(\partial J(y^\star)\big). \tag{ND}$$

Suppose $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}^R_{x^\star}), J \in \mathrm{PSF}_{y^\star}(\mathcal{M}^J_{y^\star})$ are partly smooth, denote $T^R_{x^\star}, T^J_{y^\star}$ the tangent spaces of $\mathcal{M}^R_{x^\star}, \mathcal{M}^J_{y^\star}$ at $x^\star, y^\star$, respectively. Define the following smooth perturbation of R, J,

$$\overline{R}(x) \stackrel{\text{def}}{=} \frac{1}{\gamma}\big(R(x) - \langle x, -A^T\psi^\star\rangle\big), \quad \overline{J}(y) \stackrel{\text{def}}{=} \frac{1}{\gamma}\big(J(y) - \langle w, -B^T\psi^\star\rangle\big), \tag{B.3}$$

their Riemannian Hessian $H_{\overline{R}} \stackrel{\text{def}}{=} \mathcal{P}_{T^R_{x^\star}} \nabla^2_{\mathcal{M}^R_{x^\star}} \overline{R}(x^\star) \mathcal{P}_{T^R_{x^\star}}$, $H_{\overline{J}} \stackrel{\text{def}}{=} \mathcal{P}_{T^J_{x^\star}} \nabla^2_{\mathcal{M}^J_{y^\star}} \overline{J}(y^\star) \mathcal{P}_{T^J_{x^\star}}$ and

$$\begin{aligned} M_{\overline{R}} &\stackrel{\text{def}}{=} A_R\big(\mathrm{Id} + (A_R^TA_R)^{-1}H_{\overline{R}}\big)^{-1}(A_R^TA_R)^{-1}A_R^T, \\ M_{\overline{J}} &\stackrel{\text{def}}{=} B_J\big(\mathrm{Id} + (B_J^TB_J)^{-1}H_{\overline{J}}\big)^{-1}(B_J^TB_J)^{-1}B_J^T, \end{aligned} \tag{B.4}$$

where $A_R \stackrel{\text{def}}{=} A \circ \mathcal{P}_{T^R_{x^\star}}, B_J \stackrel{\text{def}}{=} B \circ \mathcal{P}_{T^J_{y^\star}}$. Finally, define

$$M_{\mathrm{ADMM}} \stackrel{\text{def}}{=} \frac{1}{2}\mathrm{Id} + \frac{1}{2}(2M_{\overline{R}} - \mathrm{Id})(2M_{\overline{J}} - \mathrm{Id}). \tag{B.5}$$

Proof of Theorem 2.2. The proof of Theorem 2.2 is split into several steps: finite manifold identification of ADMM, local linearisation based on partial smoothness, spectral properties of the linearised matrix, and the trajectory of $\{z_k\}_{k\in\mathbb{N}}$. Let $(x^\star, y^\star, \psi^\star)$ be a saddle-point of $\mathcal{L}(x, y; \psi)$.

1. Finite manifold identification of ADMM The finite manifold identification of ADMM is already discussed in [28], below we present a short discussion for the sake of self-consistency. At convergence of ADMM, owing to (1.2) we have

$$A^T\psi^\star = \gamma A^T\big(Ax^\star - \tfrac{1}{\gamma}(z^\star - 2\psi^\star)\big) \quad \text{and} \quad B^T\psi^\star \quad = \gamma B^T\big(By^\star - \tfrac{1}{\gamma}(z^\star - \gamma b)\big).$$

From the update of x_k, y_k in (1.2), we have the following monotone inclusions

$$\begin{aligned} -\gamma A^T\big(Ax_k - \tfrac{1}{\gamma}(z_{k-1} - 2\psi_{k-1})\big) \in \partial R(x_k) \quad \text{and} \quad -\gamma B^T\big(By_k - \tfrac{1}{\gamma}(z_k - \gamma b)\big) \in \partial J(y_k), \\ -\gamma A^T\big(Ax^\star - \tfrac{1}{\gamma}(z^\star - 2\psi^\star)\big) \in \partial R(x^\star) \quad \text{and} \quad -\gamma B^T\big(By^\star - \tfrac{1}{\gamma}(z^\star - \gamma b)\big) \in \partial J(y^\star). \end{aligned}$$

Since A is bounded, it then follows that

$$\begin{aligned} \mathrm{dist}\big(-A^T\psi^\star, \partial R(x_k)\big) &\le \gamma\|A^T\big(Ax_k - \tfrac{1}{\gamma}(z_{k-1} - 2\psi_{k-1})\big) - A^T\big(Ax^\star - \tfrac{1}{\gamma}(z^\star - 2\psi^\star)\big)\| \\ &\le \gamma\|A\|\|A(x_k - x^\star) - \tfrac{1}{\gamma}(z_{k-1} - z^\star) + \tfrac{2}{\gamma}(\psi_{k-1} - \psi^\star)\| \\ &\le \gamma\|A\|\big(\|A\|\|x_k - x^\star\| + \tfrac{1}{\gamma}\|z_{k-1} - z^\star\| + \tfrac{2}{\gamma}\|\psi_{k-1} - \psi^\star\|\big) \to 0. \end{aligned}$$

and similarly

$$\mathrm{dist}\big(-B^T\psi^\star, \partial J(y_k)\big) \le \gamma\|B\|\big(\|B\|\|y_k - y^\star\| + \tfrac{1}{\gamma}\|z_k - z^\star\|\big) \to 0.$$

Since $R \in \Gamma_0(\mathbb{R}^n)$ and $J \in \Gamma_0(\mathbb{R}^m)$, then by the sub-differentially continuous property of them we have $R(x_k) \to R(x^\star)$ and $J(y_k) \to J(y^\star)$. Hence the conditions of [21, Theorem 5.3] are fulfilled for R and J, and there exists K large enough such that for all $k \ge K$, there holds

$$(x_k, y_k) \in \mathcal{M}^R_{x^\star} \times \mathcal{M}^J_{y^\star},$$

which is the finite manifold identification.

2. Linearisation of ADMM For convenience, denote $\beta = 1/\gamma$. For the update of y_k, define $w_k = -\beta(z_k - \gamma b)$, we have from (1.2) that

$$y_k = \operatorname{argmin}_{y\in\mathbb{R}^m} \beta J(y) + \frac{1}{2}\|By - w_k\|^2$$

Owing to the optimality condition of a saddle point, define $\overline{J}(y) \overset{\text{def}}{=} \beta J(y) - \langle y, -\beta B^T\psi^\star\rangle$ and its Riemannian Hessian $H_{\overline{J}} = \mathcal{P}_{T^J_{y^\star}} \nabla^2_{\mathcal{M}^J_{y^\star}} \overline{J}(y^\star)\mathcal{P}_{T^J_{y^\star}}$. For B, define $B_J = B \circ \mathcal{P}_{T^J_{y^\star}}$, and

$$M_{\overline{J}} = B_J(\mathrm{Id} + (B_J^T B_J)^{-1} H_{\overline{J}})^{-1}(B_J^T B_J)^{-1} B_J^T.$$

Then owing to Lemma A.9, we get

$$\begin{aligned} B_J(y_k - y_{k-1}) &= M_{\overline{J}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|) \\ &= -\beta M_{\overline{J}}(z_k - z_{k-1}) + o(\|z_k - z_{k-1}\|). \end{aligned} \tag{B.6}$$

Now we turn to x_k. Define $w_k = \beta(z_{k-1} - 2\psi_{k-1})$, then we get from (1.2) that

$$x_k = \operatorname{argmin}_{x\in\mathbb{R}^n} \beta R(x) + \frac{1}{2}\|Ax - w_k\|^2.$$

Define $\overline{R}(x) \overset{\text{def}}{=} \beta R(x) - \langle x, -\beta A^T\psi^\star\rangle$ and $H_{\overline{R}} = \mathcal{P}_{T^R_{x^\star}} \nabla^2_{\mathcal{M}^R_{x^\star}} \overline{R}(x^\star)\mathcal{P}_{T^R_{x^\star}}$. Denote $A_R = A \circ \mathcal{P}_{T^R_{x^\star}}$, and

$$M_{\overline{R}} = A_R(\mathrm{Id} + (A_R^T A_R)^{-1} H_{\overline{R}})^{-1}(A_R^T A_R)^{-1} A_R^T.$$

Note from (1.2) that $\psi_{k-1} - \psi_{k-2} = z_{k-1} - z_{k-2} + \gamma B(y_{k-1} - y_{k-2})$, then

$$\begin{aligned} w_k - w_{k-1} &= \beta(z_{k-1} - z_{k-2}) - 2\beta(\psi_{k-1} - \psi_{k-2}) \\ &= -\beta(z_{k-1} - z_{k-2}) - 2\beta\gamma B(y_{k-1} - y_{k-2}) \\ &= -\beta(z_{k-1} - z_{k-2}) - 2B_J(y_{k-1} - y_{k-2}) + o(\|y_{k-1} - y_{k-2}\|), \end{aligned}$$

where $y_{k-1} - y_{k-2} = \mathcal{P}_{T^R_{x^\star}}(y_{k-1} - y_{k-2}) + o(\|y_{k-1} - y_{k-2}\|)$ from Lemma A.4 is applied. From (B.2), we have $o(\|y_{k-1} - y_{k-2}\|) = o(\|z_{k-1} - z_{k-2}\|)$ and $o(\|w_{k-1} - w_{k-2}\|) = o(\|z_{k-1} - z_{k-2}\|)$, then applying Lemma A.9 yields,

$$\begin{aligned} A_R(x_k - x_{k-1}) &= M_{\overline{R}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|) \\ &= -\beta M_{\overline{R}}(z_{k-1} - z_{k-2}) + 2M_{\overline{R}} B_J(y_{k-1} - y_{k-2}) + o(\|z_{k-1} - z_{k-2}\|) \\ &= -\beta M_{\overline{R}}(z_{k-1} - z_{k-2}) + 2\beta M_{\overline{R}} M_{\overline{J}}(z_{k-1} - z_{k-2}) + o(\|z_{k-1} - z_{k-2}\|). \end{aligned} \tag{B.7}$$

Finally, from (1.2), (B.6) and (B.7), we have that

$$\begin{aligned} z_k - z_{k-1} &= \big(z_{k-1} + \gamma(Ax_k + By_{k-1} - b)\big) - \big(z_{k-2} + \gamma(Ax_{k-1} + By_{k-2} - b)\big) \\ &= (z_{k-1} - z_{k-2}) + \gamma A(x_k - x_{k-1}) + \gamma B(y_{k-1} - y_{k-2}) \\ &= (z_{k-1} - z_{k-2}) + \gamma A_R(x_k - x_{k-1}) + \gamma B_J(y_{k-1} - y_{k-2}) + o(\|z_{k-1} - z_{k-2}\|) \\ &= (z_{k-1} - z_{k-2}) - M_{\overline{R}}(z_{k-1} - z_{k-2}) + 2M_{\overline{R}} M_{\overline{J}}(z_{k-1} - z_{k-2}) + M_{\overline{J}}(z_{k-1} - z_{k-2}) \\ &\quad + o(\|z_{k-1} - z_{k-2}\|) \\ &= \big(\mathrm{Id} + 2M_{\overline{R}} M_{\overline{J}} - M_{\overline{R}} - M_{\overline{J}}\big)(z_{k-1} - z_{k-2}) + o(\|z_{k-1} - z_{k-2}\|), \end{aligned}$$

which is the desired linearisation of ADMM.

3. Spectral properties of M_{ADMM} Consider first the case where both R, J are general partly smooth functions, under which we can shown the non-expansiveness of M_{ADMM}. For $M_{\overline{R}}$, since A is injective, so is A_R, then $A_R^T A_R$ is symmetric positive definite. Therefore, we have the following similarity result for $M_{\overline{R}}$,

$$\begin{aligned} M_{\overline{R}} &= A_R\Big((A_R^T A_R)^{-\frac{1}{2}}\big(\mathrm{Id} + (A_R^T A_R)^{-\frac{1}{2}} H_{\overline{R}}(A_R^T A_R)^{-\frac{1}{2}}\big)(A_R^T A_R)^{\frac{1}{2}}\Big)^{-1}(A_R^T A_R)^{-1} A_R^T \\ &= A_R(A_R^T A_R)^{-\frac{1}{2}}\big(\mathrm{Id} + (A_R^T A_R)^{-\frac{1}{2}} H_{\overline{R}}(A_R^T A_R)^{-\frac{1}{2}}\big)^{-1}(A_R^T A_R)^{\frac{1}{2}}(A_R^T A_R)^{-1} A_R^T \\ &= A_R(A_R^T A_R)^{-\frac{1}{2}}\big(\mathrm{Id} + (A_R^T A_R)^{-\frac{1}{2}} H_{\overline{R}}(A_R^T A_R)^{-\frac{1}{2}}\big)^{-1}(A_R^T A_R)^{-\frac{1}{2}} A_R^T. \end{aligned} \tag{B.8}$$

Since $(A_R^TA_R)^{-\frac{1}{2}}H_{\overline{R}}(A_R^TA_R)^{-\frac{1}{2}}$ is symmetric positive definite, hence maximal monotone, then the matrix

$$(\mathrm{Id}+(A_R^TA_R)^{-\frac{1}{2}}H_{\overline{R}}(A_R^TA_R)^{-\frac{1}{2}})^{-1}$$

is firmly non-expansive. Let $A_R=USV^T$ be the SVD of A_R, then we have

$$\|A_R(A_R^TA_R)^{-\frac{1}{2}}\|=\|USV^T(VSU^TUSV^T)^{-\frac{1}{2}}\|=\|USV^T(VS^2V^T)^{-\frac{1}{2}}\|=\|USV^TVS^{-1}V^T\|=1.$$

Then owing to [5, Example 4.14], $M_{\overline{R}}$ is firmly non-expansive. Similarly, $M_{\overline{J}}$ is firmly non-expansive, and so is M_{ADMM} [5, Proposition 4.31]. Therefore, the power M_{ADMM}^k is convergent.

Now suppose that both R, J are locally polyhedral around $(x^\star,y^\star)$, then $M_{\overline{R}}$ and $M_{\overline{J}}$ become

$$M_{\overline{R}}=A_R(A_R^TA_R)^{-1}A_R^T\quad\text{and}\quad M_{\overline{J}}=B_J(B_J^TB_J)^{-1}B_J^T,$$

which are projection operators onto the ranges of A_R and B_J respectively. Denote these two subspaces by T_{A_R} and T_{B_J}, and correspondingly $\mathcal{P}_{T_{A_R}}\overset{\text{def}}{=}A_R(A_R^TA_R)^{-1}A_R^T$ and $\mathcal{P}_{T_{B_J}}\overset{\text{def}}{=}B_J(B_J^TB_J)^{-1}B_J^T$. Then

$$M_{\mathrm{ADMM}}=\mathcal{P}_{T_{A_R}}\mathcal{P}_{T_{B_J}}+(\mathrm{Id}-\mathcal{P}_{T_{A_R}})(\mathrm{Id}-\mathcal{P}_{T_{B_J}}).$$

Denote the dimension of T_{A_R},T_{B_J} by $\dim(T_{A_R})=p,\dim(T_{B_J})=q$, and the dimension of the intersection $\dim(T_{A_R}\cap T_{B_J})=d$. Without the loss of generality, we assume that $1\leq p\leq q\leq n$. Consequently, there are $r=p-d$ principal angles $(\zeta_i)_{i=1,\ldots,r}$ between T_{A_R} and T_{B_J} that are strictly greater than 0 and smaller than $\pi/2$. Suppose that $\zeta_1\leq\cdots\leq\zeta_r$. Define the following two diagonal matrices

$$C=\mathrm{diag}\big(\cos(\zeta_1),\cdots,\cos(\zeta_r)\big)\quad\text{and}\quad S=\mathrm{diag}\big(\sin(\zeta_1),\cdots,\sin(\zeta_r)\big).$$

Owing to [7, 16], there exists a real orthogonal matrix U such that

$$M_{\mathrm{ADMM}}=U\left[\begin{array}{cc|cc}C^2&CS&0&0\\-CS&C^2&0&0\\\hline 0&0&0_{q-p+2d}&0\\0&0&0&\mathrm{Id}_{n-p-q}\end{array}\right]U^T,$$

which indicates M_{ADMM} is normal and all its eigenvalues are inside unit disc.

Let $M_{\mathrm{ADMM}}^\infty=\lim_{k\to+\infty}M_{\mathrm{ADMM}}^k$ and $\widetilde{M}_{\mathrm{ADMM}}=M_{\mathrm{ADMM}}-M_{\mathrm{ADMM}}^\infty$, then we have

$$\widetilde{M}_{\mathrm{ADMM}}=U\left[\begin{array}{cc|c}C^2&CS&0\\-CS&C^2&0\\\hline 0&0&0_{n-2r}\end{array}\right]U^T.\tag{B.9}$$

4. Trajectory of ADMM Owing to the polyhedrality of R and J, all the small o-terms in the linearisation proof vanish and we get directly

$$z_k-z_{k-1}=M_{\mathrm{ADMM}}(z_{k-1}-z_{k-2})=M_{\mathrm{ADMM}}^k(z_0-z_{-1}).\tag{B.10}$$

As $v_k\overset{\text{def}}{=}z_k-z_{k-1}\to 0$, passing to the limit we get from above

$$0=\lim_{k\to+\infty}M_{\mathrm{ADMM}}^kv_0=M_{\mathrm{ADMM}}^\infty v_0,$$

which means $v_0\in\ker(M_{\mathrm{ADMM}})$ where $\ker(M_{\mathrm{ADMM}})$ denotes the kernel of M_{ADMM}. Since $M_{\mathrm{ADMM}}^\infty M_{\mathrm{ADMM}}^k=M_{\mathrm{ADMM}}^\infty$, we have $v_k\in\ker(M_{\mathrm{ADMM}})$ holds for any $k\in\mathbb{N}$. Then from (B.10) we have

$$v_k=(M_{\mathrm{ADMM}}-M_{\mathrm{ADMM}}^\infty)v_k=\widetilde{M}_{\mathrm{ADMM}}v_{k-1}.$$

The block diagonal property of (B.9) indicates that there exists an elementary transformation matrix E such that

$$\widetilde{M}_{\mathrm{ADMM}}=UE\left[\begin{array}{ccc}B_1&&\\&\ddots&\\&&B_r\\&&&0_{n-2r}\end{array}\right]EU^T,$$

where for each $i = 1, ..., r$, we have

$$B_i = \cos(\zeta_i) \begin{bmatrix} \cos(\zeta_i) & \sin(\zeta_i) \\ -\sin(\zeta_i) & \cos(\zeta_i) \end{bmatrix}$$

which is rotation matrix scaled by $\cos(\zeta_i)$. It is easy to show that, for each $i = 1, ..., d$, there holds

$$\lim_{k \to +\infty} B_i^k = 0,$$

since the spectral radius of B_i is $\rho(B_i) = \cos(\zeta_i) < 1$.

Suppose for some $1 \leq e < r$, we have

$$\zeta = \zeta_1 = \cdots = \zeta_e < \zeta_{e+1} \leq \cdots \leq \zeta_r.$$

Consider the following decompositions

$$\Gamma_1 = \begin{bmatrix} B_1 & & & \\ & \ddots & & \\ & & B_e & \\ & & & 0_{n-2e} \end{bmatrix} \quad \text{and} \quad \Gamma_2 = \begin{bmatrix} B_1 & & & \\ & \ddots & & \\ & & B_r & \\ & & & 0_{n-2r} \end{bmatrix} - \Gamma_1.$$

Denote $\eta = \frac{\cos(\zeta_{e+1})}{\cos(\zeta)}$, it is immediate to see that $\frac{1}{\cos^k(\zeta)} \Gamma_2^k = O(\eta^k) \to 0$, and for each $i = 1, ..., e$

$$\frac{1}{\cos(\zeta)} B_i = \begin{bmatrix} \cos(\zeta) & \sin(\zeta) \\ -\sin(\zeta) & \cos(\zeta) \end{bmatrix}$$

which is a circular rotation. Therefore, $\frac{1}{\cos(\zeta)} \Gamma_1$ is a rotation with respect to the first $2e$ elements. Denote $u_k = EU^T v_k$, then from $v_k = \widetilde{M} v_{k-1} = UE(\Gamma_1 + \Gamma_2) EU^T v_k$, we get

$$u_k = (\Gamma_1 + \Gamma_2) u_k = (\Gamma_1 + \Gamma_2)^k u_0 = \Gamma_1^k u_0 + \Gamma_2^k u_0,$$

which is an orthogonal decomposition of u_k. Define

$$s_k = \tfrac{1}{\cos^k(\zeta)} \Gamma_1^k u_1 \quad \text{and} \quad t_k = \tfrac{1}{\cos^k(\zeta)} \Gamma_2^k u_1,$$

then we have that $\|s_k\| = \|s_{k-1}\|$ and $\langle s_k,\, s_{k-1} \rangle = \cos(\zeta) \|s_k\|^2$, and $t_k = O(\eta^k)$. As a result, for $\cos(\theta_k)$ we have

$$\begin{aligned} \cos(\theta_k) = \frac{\langle v_k,\, v_{k-1} \rangle}{\|v_k\| \|v_{k-1}\|} = \frac{\langle u_k,\, u_{k-1} \rangle}{\|u_k\| \|u_{k-1}\|} &= \frac{\langle s_k + t_k,\, s_{k-1} + t_{k-1} \rangle}{\|s_k + t_k\| \|s_{k-1} + t_{k-1}\|} \\ &= \frac{\langle s_k,\, s_{k-1} \rangle}{\|s_k + t_k\| \|s_{k-1} + t_{k-1}\|} + \frac{\langle t_k,\, t_{k-1} \rangle}{\|s_k + t_k\| \|s_{k-1} + t_{k-1}\|} \\ &= \frac{\|s_k\|^2 \cos(\zeta)}{\|s_k\|^2 + \|t_k\|^2} \cdot \frac{\|s_k + t_k\|}{\|s_{k-1} + t_{k-1}\|} + O(\eta^{2k-1}). \end{aligned} \tag{B.11}$$

Using the fact that

$$\frac{\|s_k\|^2 \cos(\zeta)}{\|s_k\|^2 + \|t_k\|^2} = \cos(\zeta)\big(1 - \|t_k\|^2 + O(\|t_k\|^4)\big) = \cos(\zeta) + O(\eta^{2k}) \quad \text{and} \quad \frac{\|s_k + t_k\|}{\|s_{k-1} + t_{k-1}\|} \to 1$$

we conclude that $\cos(\theta_k) \to \cos(\zeta)$. As a matter of fact, we have $\cos(\theta_k) - \cos(\zeta) = O(\eta^{2k})$ which shows how fast $\cos(\theta_k)$ converges to $\cos(\zeta)$. □

B.3 Trajectory of ADMM: R or/and J is smooth

Now we consider the case that at least one function out of R, J is smooth. For simplicity, consider that R is smooth and J remains non-smooth. Assume that R is locally C^2-smooth around $x^\star$, the Hessian of R at $x^\star$ reads $\nabla^2 R(x^\star)$ which is positive semi-definite owing to convexity. Define $M_R \overset{\text{def}}{=} A\big(\mathrm{Id} + \frac{1}{\gamma}(A^T A)^{-1} \nabla^2 R(x^\star)\big)^{-1} (A^T A)^{-1} A^T$, and redefine

$$M_{\text{ADMM}} \overset{\text{def}}{=} \tfrac{1}{2}\mathrm{Id} + \tfrac{1}{2}(2M_R - \mathrm{Id})(2M_{\bar{J}} - \mathrm{Id}). \tag{B.12}$$

Proof of Proposition 2.4. We prove the corollary in two steps.

1. Linearisation of ADMM Following the above proof, we have for y_k that

$$B_J(y_k - y_{k-1}) = \beta M_{\bar{J}}(z_k - z_{k-1}) + o(\|z_k - z_{k-1}\|).$$

From (1.2), for x_{k+1} and x_k, since R is globally smooth differentiable

$$-A^T\big(Ax_k - \beta(z_{k-1} - 2\psi_{k-1})\big) \in \beta\nabla R(x_k) \text{ and } -A^T\big(Ax_{k-1} - \beta(z_{k-2} - 2\psi_{k-2})\big) \in \beta\nabla R(x_{k-1}),$$

which leads to, applying the local C^2-smoothness of R around $x^\star$

$$\begin{aligned}
&- A^T\big(Ax_k - \beta(z_{k-1} - 2\psi_{k-1})\big) + A^T\big(Ax_{k-1} - \beta(z_{k-2} - 2\psi_{k-2})\big)\\
&= \beta\nabla R(x_k) - \beta\nabla R(x_{k-1})\\
&= \beta\nabla^2 R(x_{k-1})(x_k - x_{k-1}) + o(\|x_k - x_{k-1}\|)\\
&= \beta\nabla^2 R(x^\star)(x_k - x_{k-1}) + \beta\big(\nabla^2 R(x_{k-1}) - \nabla^2 R(x^\star)\big)(x_k - x_{k-1}) + o(\|x_k - x_{k-1}\|)\\
&= \beta\nabla^2 R(x^\star)(x_k - x_{k-1}) + o(\|z_{k-1} - z_{k-2}\|).
\end{aligned}$$

Using the fact that A^TA is invertible and rearranging terms, we arrive at

$$\begin{aligned}
&\big(\mathrm{Id} + \beta(A^TA)^{-1}\nabla^2 R(x^\star)\big)(x_k - x_{k-1}) + o(\|z_{k-1} - z_{k-2}\|)\\
&= \beta(A^TA)^{-1}A^T(z_{k-1} - z_{k-2}) - 2\beta(A^TA)^{-1}A^T(\psi_{k-1} - \psi_{k-2}) + o(\|z_{k-1} - z_{k-2}\|)\\
&= -\beta(A^TA)^{-1}A^T(z_{k-1} - z_{k-2}) + 2(A^TA)^{-1}A^TB_J(y_{k-1} - y_{k-2}) + o(\|z_{k-1} - z_{k-2}\|),
\end{aligned}$$

which further leads to, denote $M_R = A(\mathrm{Id} + (A^TA)^{-1}H_R)^{-1}(A^TA)^{-1}A^T$

$$\begin{aligned}
A(x_k - x_{k-1}) &= -\beta M_R(z_{k-1} - z_{k-2}) + 2M_RB_J(y_{k-1} - y_{k-2}) + o(\|z_{k-1} - z_{k-2}\|)\\
&= -\beta M_R(z_{k-1} - z_{k-2}) + 2\beta M_RM_{\bar{J}}(z_{k-1} - z_{k-2}) + o(\|z_{k-1} - z_{k-2}\|).
\end{aligned}$$

Finally, from (1.2), we have that

$$z_k - z_{k-1} = \big(\mathrm{Id} + 2M_RM_{\bar{J}} - M_R - M_{\bar{J}}\big)(z_{k-1} - z_{k-2}) + o(\|z_{k-1} - z_{k-2}\|).$$

2. Trajectory of ADMM Since A is full rank square matrix and hence invertible, from (B.8) we have

$$\begin{aligned}
M_R &= A(\mathrm{Id} + \tfrac{1}{\gamma}(A^TA)^{-1}\nabla^2 R(x^\star))^{-1}(A^TA)^{-1}A^T\\
&= A(A^TA)^{-\frac{1}{2}}\big(\mathrm{Id} + \tfrac{1}{\gamma}(A^TA)^{-\frac{1}{2}}\nabla^2 R(x^\star)(A^TA)^{-\frac{1}{2}}\big)^{-1}(A^TA)^{-\frac{1}{2}}A^T\\
&\sim \big(\mathrm{Id} + \tfrac{1}{\gamma}(A^TA)^{-\frac{1}{2}}\nabla^2 R(x^\star)(A^TA)^{-\frac{1}{2}}\big)^{-1},
\end{aligned}$$

where $\big(\mathrm{Id} + \frac{1}{\gamma}(A^TA)^{-\frac{1}{2}}\nabla^2 R(x^\star)(A^TA)^{-\frac{1}{2}}\big)^{-1}$ is symmetric positive definite. If we choose γ such that

$$\tfrac{1}{\gamma}\|(A^TA)^{-\frac{1}{2}}\nabla^2 R(x^\star)(A^TA)^{-\frac{1}{2}}\| < 1,$$

then all the eigenvalues of M_R are in $]1/2, 1]$, hence $W_R \overset{\text{def}}{=} 2M_R - \mathrm{Id}$ is symmetric positive definite. Therefore, we get

$$\begin{aligned}
\tfrac{1}{2}\mathrm{Id} + \tfrac{1}{2}W_R\big(2M_{\bar{J}} - \mathrm{Id}\big) &= W_R^{1/2}\big(\tfrac{1}{2}\mathrm{Id} + \tfrac{1}{2}W_R^{1/2}\big(2M_{\bar{J}} - \mathrm{Id}\big)W_R^{1/2}\big)W_R^{-1/2}\\
&\sim \tfrac{1}{2}\mathrm{Id} + \tfrac{1}{2}W_R^{1/2}\big(2M_{\bar{J}} - \mathrm{Id}\big)W_R^{1/2},
\end{aligned}$$

and $\overline{M} \overset{\text{def}}{=} \frac{1}{2}\mathrm{Id} + \frac{1}{2}W_R^{1/2}(2M_{\bar{J}} - \mathrm{Id})W_R^{1/2}$ is symmetric positive semi-definite with all eigenvalues in $[0, 1]$. Hence, by similarity, the eigenvalues of M are all real and contained in $[0, 1]$. □

C Adaptive acceleration for ADMM

C.1 Convergence of A^{3}DMM

Proof of Proposition 4.2. From (4.3), we have that

$$z_k = \mathcal{F}(z_{k-1} + \varepsilon_{k-1}) = \mathcal{F}(z_{k-1}) + \big(\mathcal{F}(z_{k-1} + \varepsilon_{k-1}) - \mathcal{F}(z_{k-1})\big).$$

Given any $z^\star \in \mathrm{fix}(\mathcal{F})$, since $\mathcal{F}$ is firmly non-expansive, hence non-expansive, we have

$$\|z_k - z^\star\| \le \|\mathcal{F}(z_{k-1}) - \mathcal{F}(z^\star)\| + \|\mathcal{F}(z_{k-1} + \varepsilon_k) - \mathcal{F}(z_{k-1})\| \le \|z_{k-1} - z^\star\| + \|\varepsilon_{k-1}\|,$$

which means that $\{z_k\}_{k\in\mathbb{N}}$ is quasi-Fejér monotone with respect to $\mathrm{fix}(\mathcal{F})$. Then invoke [5, Proposition 5.34] we obtain the convergence of the sequence $\{z_k\}_{k\in\mathbb{N}}$. □

C.2 Acceleration guarantee of A^{3}DMM

Recall the definition of V_{k-1}, c_k, C_k and $\bar{z}_{k,s}$ in the beginning of the section. By definition,

$$V_k = MV_{k-1}. \tag{C.1}$$

Define $E_{k,j} \overset{\text{def}}{=} V_k C_k^j - V_{k+1}$ for $j \ge 1$ and

$$E_{k,0} \overset{\text{def}}{=} V_{k-1}C_k - V_k = \begin{bmatrix}(V_{k-1}c_k - v_k) & 0 & \cdots & 0\end{bmatrix}. \tag{C.2}$$

We obtain the relation between the extrapolated point $\bar{z}_{k,s}$ and the $(k+s)$'th point of $\{z_k\}_{k\in\mathbb{N}}$

$$\bar{z}_{k,s} = z_k + \textstyle\sum_{j=1}^{s}(v_{j+k} + (E_{k,j})_{(:,1)}) = z_{k+s} + \sum_{j=1}^{s}(E_{k,j})_{(:,1)}$$

In the following, given a matrix M, we let $\rho(M)$ denote the spectral radius of M and $\lambda(M)$ denote its spectrum.

Proof of Proposition 4.3. We first prove (i) that the extrapolation error is controlled by the coefficients fitting error. Since $k \in \mathbb{N}$ is fixed, for ease of notation, we also write $E_\ell = E_{k,\ell}$ and $C = C_k$. We first show that for $\ell \in \mathbb{N}$, we have

$$E_\ell = \textstyle\sum_{j=1}^{\ell} M^j E_0 C^{\ell-j}. \tag{C.3}$$

We prove this by induction. Note that

$$V_kC \overset{(C.1)}{=} (MV_{k-1})C \overset{(C.2)}{=} MV_k + ME_0 \overset{(C.1)}{=} V_{k+1} + ME_0.$$

Therefore, $E_1 = ME_0$ as required. Assume that (C.4) is true up to $\ell = m$. Then,

$$\begin{aligned} V_kC^{m+1} &\overset{(C.1)}{=} (MV_{k-1})C^{m+1} \overset{(C.2)}{=} MV_kC^m + ME_0C^m = M(V_{m+k} + E_m) + ME_0C^m \\ &\overset{(C.1)}{=} V_{m+2} + ME_m + ME_0C^m \end{aligned}$$

So, plugging in our assumption on E_m, we have

$$E_{m+1} = ME_m + ME_0C^m = ME_0C^m + M\big(\textstyle\sum_{j=1}^{m} M^j E_0 C^{m-j}\big) = \sum_{j=1}^{m+1} M^j E_0 C^{m+1-j}.$$

To bound the extrapolation error,

$$\textstyle\sum_{m=1}^{s} E_m = \sum_{m=1}^{s}(\sum_{j=1}^{m} M^j E_0 C^{m-j}) = \sum_{\ell=0}^{s-1}\big(\sum_{j=1}^{s-\ell} M^j\big) E_0 C^\ell = \sum_{\ell=1}^{s} M^\ell E_0 \big(\sum_{i=0}^{s-\ell} C^i\big)$$

Therefore,

$$\|\bar{z}_{k,s} - z^\star\| \le \|z_{k+s} - z^\star\| + \textstyle\sum_{\ell=1}^{s} \|M^\ell\| \|E_0\| \| \sum_{i=0}^{s-\ell} C^i_{(1,1)}\|.$$

In the case of $s = +\infty$, we have

$$\|\bar{z}_{k,\infty} - z^\star\| \leq \sum_{\ell=1}^{\infty} \|M^\ell\| \|E_0 (\mathrm{Id} - C)^{-1}_{(:,1)}\| = \frac{\|E_0\|}{1 - \sum_i c_i} \sum_{\ell=1}^{\infty} \|M^\ell\|.$$

The fact that B_s is uniformly bounded in s if $\rho(M) < 1$ and $\rho(C) < 1$ follows because this implies that $\sum_{\ell=1}^{\infty} \|M^\ell\| < \infty$ thanks to the Gelfand formula, and $\sum_{i=0}^{\infty} C^i = (\mathrm{Id} - C)^{-1}$ and its $(1,1)^{th}$ entry is precisely $\frac{1}{1-\sum_i c_i}$. Since $k \in \mathbb{N}$ is fixed, for ease of notation, we also write $E_\ell = E_{k,\ell}$ and $C = C_k$. We first show that for $\ell \in \mathbb{N}$, we have

$$E_\ell = \sum_{j=1}^{\ell} M^j E_0 C^{\ell - j}. \tag{C.4}$$

We prove this by induction. Note that

$$V_k C \overset{(C.1)}{=} (M V_{k-1}) C \overset{(C.2)}{=} M V_k + M E_0 \overset{(C.1)}{=} V_{k+1} + M E_0.$$

Therefore, $E_1 = M E_0$ as required. Assume that (C.4) is true up to $\ell = m$. Then,

$$\begin{aligned} V_k C^{m+1} &\overset{(C.1)}{=} (M V_{k-1}) C^{m+1} \\ &\overset{(C.2)}{=} M V_k C^m + M E_0 C^m = M(V_{m+k} + E_m) + M E_0 C^m \\ &\overset{(C.1)}{=} V_{m+2} + M E_m + M E_0 C^m. \end{aligned}$$

So, plugging in our assumption on E_m, we have

$$E_{m+1} = M E_m + M E_0 C^m = M E_0 C^m + M\big(\sum_{j=1}^{m} M^j E_0 C^{m-j}\big) = \sum_{j=1}^{m+1} M^j E_0 C^{m+1-j}.$$

To bound the extrapolation error,

$$\sum_{m=1}^{s} E_m = \sum_{m=1}^{s} \big(\sum_{j=1}^{m} M^j E_0 C^{m-j}\big) = \sum_{\ell=0}^{s-1} \big(\sum_{j=1}^{s-\ell} M^j\big) E_0 C^\ell = \sum_{\ell=1}^{s} M^\ell E_0 \big(\sum_{i=0}^{s-\ell} C^i\big)$$

Therefore,

$$\|\bar{z}_{k,s} - z^\star\| \leq \|z_{k+s} - z^\star\| + \sum_{\ell=1}^{s} \|M^\ell\| \|E_0\| \big\| \sum_{i=0}^{s-\ell} C^i_{(1,1)} \big\|.$$

In the case of $s = +\infty$, we have

$$\|\bar{z}_{k,\infty} - z^\star\| \leq \sum_{\ell=1}^{\infty} \|M^\ell\| \|E_0 (\mathrm{Id} - C)^{-1}_{(:,1)}\| = \frac{\|E_0\|}{1 - \sum_i c_i} \sum_{\ell=1}^{\infty} \|M^\ell\|.$$

The fact that B_s is uniformly bounded in s if $\rho(M) < 1$ and $\rho(C) < 1$ follows because this implies that $\sum_{\ell=1}^{\infty} \|M^\ell\| < \infty$ thanks to the Gelfand formula, and $\sum_{i=0}^{\infty} C^i = (\mathrm{Id} - C)^{-1}$ and its $(1,1)^{th}$ entry is precisely $\frac{1}{1-\sum_i c_i}$.

To control the coefficients fitting error ϵ_k, we follow closely the arguments of Section 6.7 in [37], since this amounts to understanding the behaviour of the coefficients c_k, which are precisely the MPE coefficients. Recall our assumption that M is diagonalisable, so $M = U^\top \Sigma U$ where U is an orthogonal matrix and Σ is a diagonal matrix with the eigenvalues of M as its diagonal. Then, letting $u_k \overset{\text{def}}{=} U v_k$,

$$\begin{aligned} \epsilon_k &= \min_{c \in \mathbb{R}^q} \| \sum_{i=1}^{q} c_i v_{k-i} - v_k \| \\ &= \min_{c \in \mathbb{R}^q} \| \sum_{i=1}^{q} c_i \Sigma^{k-i} u_0 - \Sigma^k u_0 \| = \min_{g \in \mathcal{P}_q} \|\Sigma^{k-q} g(\Sigma) u_0\| \leq \|u_0\| \min_{g \in \mathcal{P}_q} \max_{z \in \lambda(M)} |z|^{k-q} |g(z)| \end{aligned}$$

where $\mathcal{P}_q$ is the set of monic polynomials of degree q and $\lambda(M)$ is the spectrum of M. Choosing $g = \prod_{j=1}^{q} (z - \lambda_j)$, we have $g(\lambda_j) = 0$ for $j = 1, \ldots, q$, so

$$\epsilon_k \leq \|u_0\| |\lambda_{q+1}|^{k-q} \max_{\ell > q} \prod_{j=1}^{q} |\lambda_j - \lambda_\ell|. \tag{C.5}$$

The claim that $\rho(C_k) < 1$ holds since the eigenvalues of C are precisely the roots of the polynomial $Q(z) = z^{k-1} - \sum_{i=1}^{k-1} c_j z^{k-1-i}$, and from [37], if $|\lambda_q| > |\lambda_{q+1}|$, then Q has precisely q roots $r_1, \dots, r_q$ satisfying $r_j = \lambda_j + \mathcal{O}(|\lambda_{q+1}/\lambda_j|^k)$. So, $|r_j| < 1$ for all k sufficiently large. To prove the non-asymptotic bounds on ϵ_k, first observe that $z_{k+1} - z_k = M(z_k - z_{k-1})$ implies $z_{k+1} - z^\star = M(z_k - z_*)$ and $z_{k+1} - z_k = (M - \mathrm{Id})(z_k - z^\star)$. So, letting $\gamma_i = -c_{k,i}/(1 - \sum_i c_{k,i})$ for $i = 1, \dots, q$ and $\gamma_0 = 1/(1 - \sum_i c_{k,i})$, we have

$$\frac{1}{1 - \sum_i c_{k,i}}\big(v_k - \textstyle\sum_{i=1}^q c_{k,i} v_{k-i}\big) = \sum_{i=0}^q \gamma_i v_{k-i} = (M - \mathrm{Id}) \sum_{i=0}^q \gamma_i (z_{k-i-1} - z^\star). \tag{C.6}$$

Now, $y \stackrel{\text{def}}{=} \sum_{i=0}^q \gamma_i z_{k-i-1}$ is precisely the MPE update and norm bounds on this are presented in [37]. For completeness, we reproduce their arguments here: Let $A \stackrel{\text{def}}{=} \mathrm{Id} - M$, by our assumption of $\lambda(M) \subset (-1, 1)$, we have that A is positive definite. Then,

$$\begin{aligned}\|A^{1/2}(y - z^\star)\|^2 &= \langle A(y - z^\star),\, (y - z^\star)\rangle \\ &= -\langle \textstyle\sum_{i=0}^q \gamma_i v_{k-i},\, (y - z^\star) + w\rangle\end{aligned}$$

where $w = \sum_{j=1}^q a_j v_{k-j}$ with $a \in \mathbb{R}^q$ being arbitrary, since by definition of γ, $\langle \sum_{i=0}^q \gamma_i v_{k-i},\, v_\ell\rangle = 0$ for all $\ell = k - q, \dots, k - 1$. We can write

$$w = \sum_{j=1}^q a_j (M - \mathrm{Id})(z_{k-j-1} - z^\star) = \sum_{j=1}^q a_j (M - \mathrm{Id}) M^{k-j-1}(z_0 - z^\star) = f(M)(z_0 - z^\star)$$

where $f(z) = z^{k-q-1}(z - 1)\sum_{j=1}^q a_j z^{q-j}$, and we can write

$$y - z^\star = \sum_{i=0}^q \gamma_i M^{k-i-1}(z_0 - z^\star) = g(M)(z_0 - z^\star)$$

where $g(z) = z^{k-q-1}\sum_{i=0}^q \gamma_i z^{q-i}$. Therefore, $f(z) + g(z) = z^{k-1-q}h(z)$, where h is a polynomial of degree q such that $h(1) = 1$. Moreover, since the coefficients a_j are arbitrary, h can be considered as an arbitrary element of $\tilde{\mathcal{P}}_q$, the set of all polynomials of degree q such that $h(1) = 1$. Therefore

$$\begin{aligned}\|A^{-1/2}(y - z^\star)\|^2 &\le \|A^{-1/2}(y - z^\star)\| \min_{h \in \tilde{\mathcal{P}}_q} \|M^n h(M)(z_0 - z^\star)\| \\ &\le \|A^{-1/2}(y - z^\star)\| \min_{h \in \tilde{\mathcal{P}}_q} \max_{t \in \lambda(M)} |t^n h(t)| \|z_0 - z^\star\|\end{aligned}$$

In particular, combining this with (C.6), we have

$$\frac{\epsilon_k}{|1 - \sum_i c_{k,i}|} \le \|z_0 - z^\star\| \|(\mathrm{Id} - M)^{1/2}\| \rho(M)^n \min_{h \in \tilde{\mathcal{P}}_q} \max_{t \in \lambda(M)} |h(t)|$$

Finally, in our case where $\lambda(M) = [\alpha, \beta]$ with $1 > \beta > \alpha > -1$, it is well known that $\min_{h \in \tilde{\mathcal{P}}_q} \max_{t \in \lambda(M)} |h(t)|$ has an explicit expression (see, for example, [9] or [37, Section 7.3.1]):

$$\min_{h \in \tilde{\mathcal{P}}_q} \max_{z \in \lambda(M)} |h(z)| \le \max_{z \in \lambda(M)} |h_*(z)|,$$

where $h_*(z) \stackrel{\text{def}}{=} \frac{T_q(\frac{2z - \alpha - \beta}{\beta - \alpha})}{T_q(\frac{2 - \alpha - \beta}{\beta - \alpha})}$ where $T_q(x)$ is the q^{th} Chebyshev polynomial and it is well known that

$$\min_{h \in \tilde{\mathcal{P}}_q} \max_{z \in [\alpha, \beta]} |h(z)| \le 2\Big(\frac{\sqrt{\eta} - 1}{\sqrt{\eta} + 1}\Big)^q \tag{C.7}$$

where $\eta = \frac{1 - \alpha}{1 - \beta}$. □