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, flexibility and efficiency. Over the years, numerous efforts are made to improve the performance of 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 analyze an adaptive acceleration scheme for the method. Numerical experiments on problems arising from image processing, statistics and machine learning demonstrate the advantages of the proposed acceleration approach.

1 Introduction

Consider the following constrained and composite optimisation problem

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

where the following basic assumptions are imposed

($\mathcal{A}$.1) $R\in\Gamma_0(\mathbb{R}^n)$ and $J\in\Gamma_0(\mathbb{R}^m)$ are proper convex and lower semi-continuous functions.

($\mathcal{A}$.2) $A:\mathbb{R}^n\to\mathbb{R}^p$ and $B:\mathbb{R}^m\to\mathbb{R}^p$ are injective linear operators.

($\mathcal{A}$.3) $\operatorname{ri}(\operatorname{dom}(R)\cap\operatorname{dom}(J))\neq\emptyset$, and the set of minimizers is non-empty.

Over the past years, problem ($\mathscr{P}_{\text{ADMM}}$) has attracted a great deal of interests as it covers many important problems arising from data science, machine learning, statistics, inverse problems and imaging science, etc.; See Section 6 for examples. In the literature, different numerical schemes are proposed to handle the problem, among them the alternating direction method of multipliers (ADMM) is the most prevailing one.

Earlier works of ADMM include [27, 26, 25, 20], and recently it has gained increasing popularity, in part due to [11]. To derive ADMM, first consider the augmented Lagrangian associated to ($\mathscr{P}_{\text{ADMM}}$) which reads

$$\mathscr{L}(x,y;\psi)\overset{\text{def}}{=}R(x)+J(y)+\langle\psi,Ax+By-b\rangle+\frac{\gamma}{2}\|Ax+By-b\|^2,$$

where $\gamma>0$ and $\psi\in\mathbb{R}^p$ is the Lagrangian multiplier. To find a saddle-point of $\mathscr{L}(x,y;\psi)$, ADMM applies the following iteration

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

By defining a new point $z_k\overset{\text{def}}{=}\psi_{k-1}+\gamma Ax_k$, we can rewrite ADMM iteration (1.1) into the following form

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

*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

For the rest of the paper, we shall focus on the above four-point formulation of ADMM. In the literature, it is well known that ADMM is equivalent to applying Douglas–Rachford splitting [19] to the dual problem of ($\mathscr{P}_{\text{ADMM}}$) [25] which reads

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

where $R^*(v)\overset{\text{def}}{=}\sup_{x\in\mathbb{R}^n}(\langle x, v\rangle - R(x))$ is the Fenchel conjugate, or simply conjugate, of R. The corresponding iteration of Douglas–Rachford splitting reads

$$\begin{aligned}
u_k &= \operatorname{argmin}_{u\in\mathbb{R}^p} \gamma R^*(-A^T u)+\tfrac{1}{2}\|u-(2\psi_{k-1}-z_{k-1})\|^2,\\
z_k &= z_{k-1}+u_k-\psi_{k-1},\\
\psi_k &= \operatorname{argmin}_{\psi\in\mathbb{R}^p} \gamma J^*(-B^T\psi)+\langle\psi,\gamma b\rangle+\tfrac{1}{2}\|\psi-z_k\|^2,
\end{aligned}$$

where z_k is exactly the same one of (1.2). The above iteration can be written as the fixed-point iteration of z_k, that is

$$z_k = \mathscr{F}_{\text{DR}}(z_{k-1}), \tag{1.3}$$

with $\mathscr{F}_{\text{DR}}$ being the fixed-point operator. We refer to Appendix C for the expression of $\mathscr{F}_{\text{DR}}$ and more discussions between the equivalence between ADMM and Douglas–Rachford splitting. Based on such an equivalence, the convergence property of ADMM has been well studied in the literature, we refer to [30, 21, 18, 31] and the references therein. In [5], u_k, ψ_k are called the "*shadow sequences*" of Douglas–Rachford splitting, in this paper, following the terminology, we shall call x_k, y_k of (1.1) the shadow sequences of ADMM.

1.1 Contributions

The contribution of our paper is threefold. First, for the sequence $\{z_k\}_{k\in\mathbb{N}}$ of (1.2), we show 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 characterized precisely if R, J are moreover locally polyhedral around the solution.
- When at least one of R, J is smooth, we show that depends on the choice of γ, the eventual trajectory of $\{z_k\}_{k\in\mathbb{N}}$ can be either straight line or spiral.

Second, based on trajectory of $\{z_k\}_{k\in\mathbb{N}}$, we discuss the limitations of the current combination of ADMM and inertial acceleration technique. In Section 3, we distinguish the situations where inertial acceleration will work and when it fails. More precisely, we find that 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 iterates. In Section 4, we propose an adaptive extrapolation 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 geometrical interpretation for polynomial extrapolation methods such as Minimal Polynomial Extrapolation (MPE) [14] and Reduced Rank Extrapolation (RRE) [22, 39].

1.2 Related works

Over the past decades, owing to the tremendous success of inertial acceleration [41, 9], the inertial technique has been widely adapted to accelerate other first-order algorithms. In terms of ADMM, related work can be found in [43, 32, 24], either from proximal point algorithm perspective or continuous dynamical system. However, to ensure that inertial acceleration works, stronger assumptions are imposed on R, J in ($\mathscr{P}_{\text{ADMM}}$), such as smooth differentiability or strong convexity. When it comes to general non-smooth problems, these works may fail to provide acceleration. Recently in [23], an $O(1/k^2)$ convergence rate is established for

ADMM using Nesterov acceleration, however the result holds only for the continuous dynamical system while the discrete-time optimization scheme remains unavailable.

For more generic acceleration techniques, there are extensive works in numerical analysis on the topic of convergence acceleration for sequences. The goal of convergence acceleration is, given an arbitrary sequence $\{z_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ with limit $z^\star$, finding a transformation $\mathscr{E}_k : \{z_{k-j}\}_{j=1}^q \to \bar{z}_k \in \mathbb{R}^n$ 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 [45], the Δ^2-process of Aitken [2] and Shank's algorithm [47]. We refer to [12, 13, 48] 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 [53] 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) [14] and Reduced Rank Extrapolation (RRE) [22, 39] (which is also a variant of Anderson acceleration developed independently in [4]), which are particularly relevant to this present work (see Section 4.2 for brief discussion).

More recently, there has been a series of work on a regularized version of RRE stemming from [46]. We remark however that the regularisation 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 inner 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 a set-valued mapping defined by $\partial R(x) \stackrel{\text{def}}{=} \left\{g \in \mathbb{R}^n | R(x') \geq R(x) + \langle g, x'-x\rangle, \forall x' \in \mathbb{R}^n\right\}$. The spectral radius of a matrix M is denoted by $\rho(M)$.

Organization The rest of paper is organized as following: in Section 2 we study the trajectory of ADMM for the sequence $\{z_k\}_{k\in\mathbb{N}}$. The limitation of the inertial technique is discussed in Section 3. Then in Section 4, we propose the adaptive acceleration scheme for ADMM followed by detailed discussions provided in Section 5. Numerical experiments from machine learning and imaging are provided in Section 6. In the appendix, we first provide extra discussions on why inertial fails, and then the proofs of main propositions.

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 [34].

2.1 Partial smoothness

The difficulty of analyzing the trajectory, or in general the behaviors (*e.g.* convergence rate), of ADMM is that the iteration of the method is non-linear. We need a proper tool of (locally) linearize the iteration of ADMM, and partial smoothness provides a powerful framework to achieve the goal. Let $\mathscr{M} \subset \mathbb{R}^n$ be a C^2-smooth Riemannian manifold, denote $\mathscr{T}_{\mathscr{M}}(x)$ the tangent space of $\mathscr{M}$ at a point $x \in \mathscr{M}$.

Definition 2.1 (Partly smooth function [34]). A function $R \in \Gamma_0(\mathbb{R}^n)$ is partly smooth at $\bar{x}$ relative to a set $\mathscr{M}_{\bar{x}}$ if $\partial R(\bar{x}) \neq \emptyset$ and $\mathscr{M}_{\bar{x}}$ is a C^2 manifold around $\bar{x}$, and moreover

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

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

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

The class of partly smooth functions at $\bar{x}$ relative to $\mathscr{M}_{\bar{x}}$ is denoted as $\mathrm{PSF}_{\bar{x}}(\mathscr{M}_{\bar{x}})$. Popular examples of partly smooth functions can be found in [35, Chapter 5]. Loosely speaking, a partly smooth function behaves *smoothly* along $\mathscr{M}_{\bar{x}}$, and *sharply* normal to it. The essence of partial smoothness is that the behaviour

of the function and of its minimizers depend essentially on its restriction to this manifold, hence providing the possibilities to study the trajectory of sequences.

2.2 Trajectory of sequence z_k

Next we discuss the trajectory of ADMM in terms of $\{z_k\}_{k\in\mathbb{N}}$. The iteration of ADMM is non-linear in general owing to the non-linearity of the proximity mappings. However, when R, J are partly smooth, the local C^2-smoothness allows us to linearize the proximity mappings, hence the ADMM iteration. In turn, this allows us to study the trajectory of sequence generated by the method. We denote $(x^\star, y^\star, \psi^\star)$ a saddle-point of $\mathscr{L}(x, y; \psi)$ and let $z^\star = \psi^\star + \gamma A x^\star$.

Denote $v_k \overset{\text{def}}{=} z_k - z_{k-1}$ and let $\theta_k \overset{\text{def}}{=} \arccos\left(\frac{\langle v_k, v_{k-1}\rangle}{\|v_k\|\|v_{k-1}\|}\right)$ be the angle between v_k, v_{k-1}. We shall use $\{\theta_k\}_{k\in\mathbb{N}}$ to characterize the trajectory of $\{z_k\}_{k\in\mathbb{N}}$. Given a saddle point $(x^\star, y^\star, \psi^\star)$, the corresponding KKT condition entails $-A^T\psi^\star \in \partial R(x^\star)$ and $-B^T\psi^\star \in \partial J(y^\star)$, below we impose that

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

2.2.1 Both R, J are non-smooth

Let $\mathscr{M}_{x^\star}^R, \mathscr{M}_{y^\star}^J$ be two smooth manifolds around $x^\star, y^\star$ respectively, and suppose $R \in \mathrm{PSF}_{x^\star}(\mathscr{M}_{x^\star}^R), J \in \mathrm{PSF}_{y^\star}(\mathscr{M}_{y^\star}^J)$ are partly smooth. Denote $T_{x^\star}^R, T_{y^\star}^J$ the tangent spaces of $\mathscr{M}_{x^\star}^R, \mathscr{M}_{y^\star}^J$ at $x^\star, y^\star$, respectively. Let $A_R \overset{\text{def}}{=} A \circ \mathcal{P}_{T_{x^\star}^R}, B_J \overset{\text{def}}{=} B \circ \mathcal{P}_{T_{y^\star}^J}$ and T_{A_R}, T_{B_J} be the range of A_R, B_J respectively. Denote $(\alpha_j)_{j=1,\ldots,\min\{\dim(T_{A_R}),\dim(T_{B_J})\}}$ the principal angles (see Definition B.1) between T_{A_R}, T_{B_J}, and let α_F, α' be the smallest and 2nd smallest of all non-zero α_j.

Theorem 2.2. *For problem ($\mathscr{P}_{\text{ADMM}}$) and ADMM iteration (1.2), 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 $\mathscr{L}(x, y; \psi)$. Suppose that $R \in \mathrm{PSF}_{x^\star}(\mathscr{M}_{x^\star}^R), J \in \mathrm{PSF}_{y^\star}(\mathscr{M}_{y^\star}^J)$ and condition (ND) holds, then*

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

(ii) *If moreover, R, J are locally polyhedral around $x^\star, y^\star$, then $v_k = M_{\text{ADMM}} v_{k-1}$ 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 functions, though we cannot prove, numerical evidence shows that the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ could be either straight line or spiral; See Section 6 for evidences.

2.2.2 R or/and J is smooth

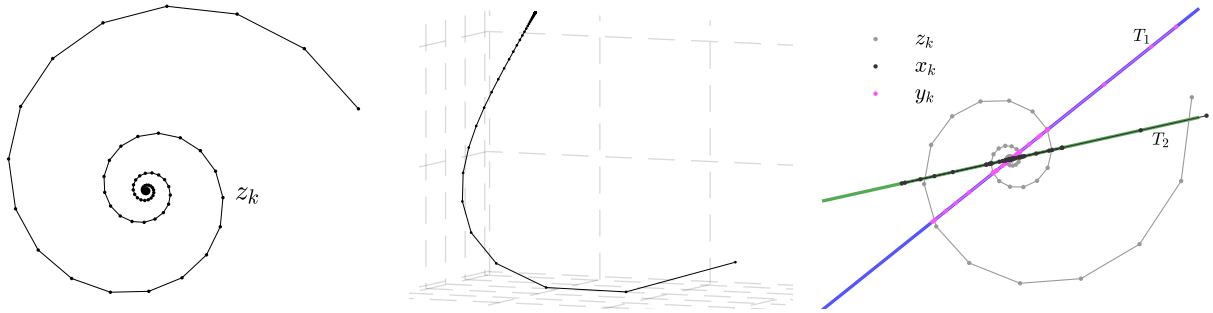
Now consider the case that at least one function out of R, J is smooth. For simplicity, consider that R is smooth and J remains non-smooth. We have the following result.

Proposition 2.4. *For problem ($\mathscr{P}_{\text{ADMM}}$) and ADMM iteration (1.2), 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 $\mathscr{L}(x, y; \psi)$. Suppose R is locally C^2 around $x^\star$, $J \in \mathrm{PSF}_{y^\star}(\mathscr{M}_{y^\star}^J)$ 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. When the spectrum of M is real, 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 moreover $o(\|v_{k-1}\|)$ vanishes fast enough, we can prove that $\theta_k \to 0$.

It should be emphasized that the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is determined by the property of the *leading eigenvalue* of M_{ADMM}. Therefore, for $\gamma \le \|(A^TA)^{-\frac{1}{2}}\nabla^2 R(x^\star)(A^TA)^{-\frac{1}{2}}\|$, though M_{ADMM} will have complex eigenvalues, the leading one is not necessarily to be complex. As a result, the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ could be either spiral (complex leading eigenvalue) or straight line (real leading eigenvalue).

In Figure 1, 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 feasibility problem; See Section 3.2. For this case, both R and J are indicator functions of affine subspaces. For subfigure (b), we use ADMM to solve a toy LASSO problem in $\mathbb{R}^3$ with γ chosen such that the eventual trajectory of $\{z_k\}_{k\in\mathbb{N}}$ is straight line.



(a) Spiral trajectory (b) Eventual straight line trajectory (c) Trajectory of shadow sequences

Figure 1: Two different trajectories of ADMM and trajectory of shadow sequences.

Remark 2.6 (Trajectory of shadow sequences x_k, y_k). We conclude this section by remarking the trajectories of the shadows sequences x_k, y_k, which are more complicated than that of z_k. The trajectories of x_k, y_k depend on both the trajectory of z_k and the properties of the functions R, J. For example, for the feasibility problem of Section 3.2, the trajectory of z_k is a spiral, see Figure 1 (a) and (c). The trajectory of y_k is the projection of the trajectory of z_k onto subspace T_1, which means that y_k is swinging around $y^\star$ along T_1. Similar trajectory for x_k; see Figure 1 (c). The trajectories of x_k and/or y_k can also be spirals when R and/or J are smooth and the trajectory of z_k is spiral. When the trajectory of z_k is a straight-line, so are the trajectories of x_k, y_k.

3 The failure of inertial acceleration

In this section, we use the LASSO and feasibility problems as examples to demonstrate the effects of applying inertial technique to ADMM, especially when it fails. One simple approach to combine inertial technique and ADMM is via the equivalence between ADMM and Douglas–Rachford splitting method. Applying the inertial scheme of [35, Chapter 4] to the Douglas–Rachford iteration (1.3), we obtain the following inertial scheme

$$\begin{aligned} \bar{z}_k &= z_k + a_k(z_k - z_{k-1}),\\ z_{k+1} &= \mathscr{F}_{\mathrm{DR}}(\bar{z}_k). \end{aligned}$$

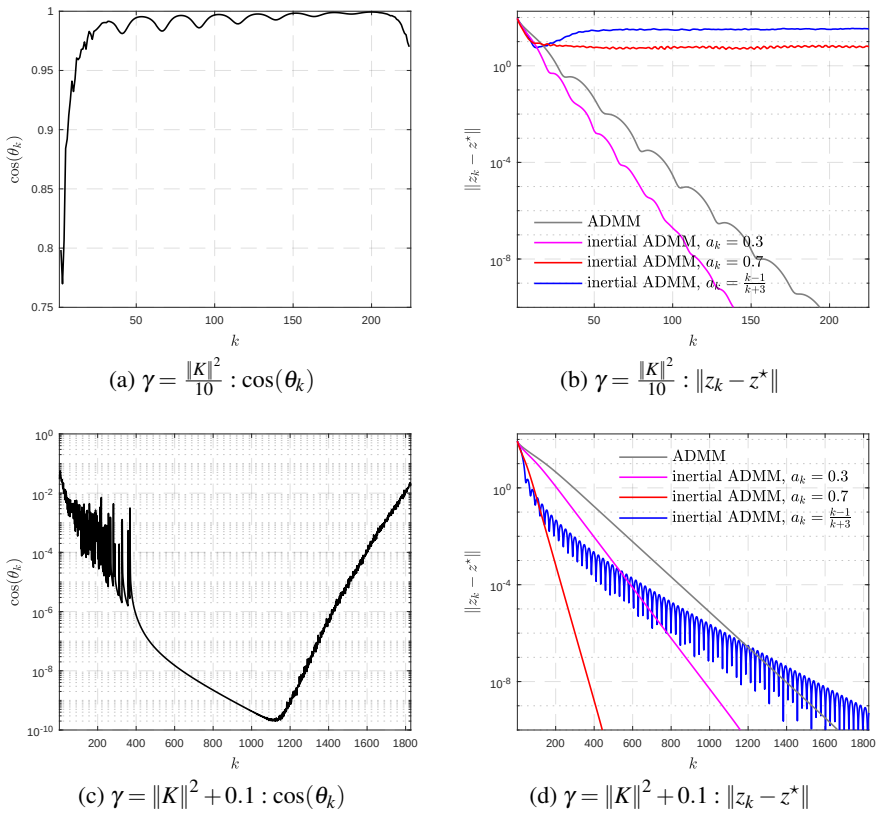
The above scheme can be reformulated as an instance of inertial Proximal Point Algorithm, guaranteed to be convergent for $a_k < \frac{1}{3}$ [3]; We refer to [43] or [35, Chapter 4.3] for more details. Adapting the above scheme to ADMM we obtain the following inertial ADMM (iADMM)

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

which considers only the momentum of $\{z_k\}_{k\in\mathbb{N}}$ without any stronger assumptions on R, J. To our knowledge, there is no acceleration guarantee for (3.1).

Remark 3.1.

- In the inertial scheme (3.1), we can also consider momentum of more than two points, that is using more points than z_k, z_{k-1} to update $\bar{z}_k$. For example, the following three-point momentum can be considered
$$\bar{z}_k = z_k + a_k(z_k - z_{k-1}) + b_k(z_{k-1} - z_{k-2}).$$
We shall use the feasibility problem to demonstrate the benefits of the above approach.
- In literature, besides (3.1), other combinations of inertial technique and ADMM are also proposed, see for instance [43, 32]. To ensure acceleration guarantees, stronger assumption needs to be imposed, such as Lipschitz smoothness and strong convexity.



(a) $\gamma = \frac{\|K\|^2}{10}$: $\cos(\theta_k)$ (b) $\gamma = \frac{\|K\|^2}{10}$: $\|z_k - z^\star\|$

(c) $\gamma = \|K\|^2 + 0.1$: $\cos(\theta_k)$ (d) $\gamma = \|K\|^2 + 0.1$: $\|z_k - z^\star\|$

Figure 2: Trajectory of sequence $\{z_k\}_{k\in\mathbb{N}}$ and performance of inertial on ADMM. (a) $\cos(\theta_k)$ for $\gamma = \frac{\|K\|^2}{10}$; (b) failure of inertial ADMM on spiral trajectory; (c) $\cos(\theta_k)$ for $\gamma = \|K\|^2 + 0.1$; (d) success of inertial ADMM on straight line trajectory.

3.1 LASSO problem

The formulation of LASSO in the form of ($\mathscr{P}_{\text{ADMM}}$) reads
$$\min_{x,y\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{2}\|Ky - f\|^2 \quad \text{such that} \quad x - y = 0, \tag{3.2}$$
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 $\gamma = \frac{\|K\|^2}{10}$ and $\gamma = \|K\|^2 + 0.1$, 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 [15]. For the numerical example, we consider $K \in \mathbb{R}^{640\times 2048}$ and $\mu = 1$, f is the measurement of an 128-sparse signal. The results are shown in Figure 2,

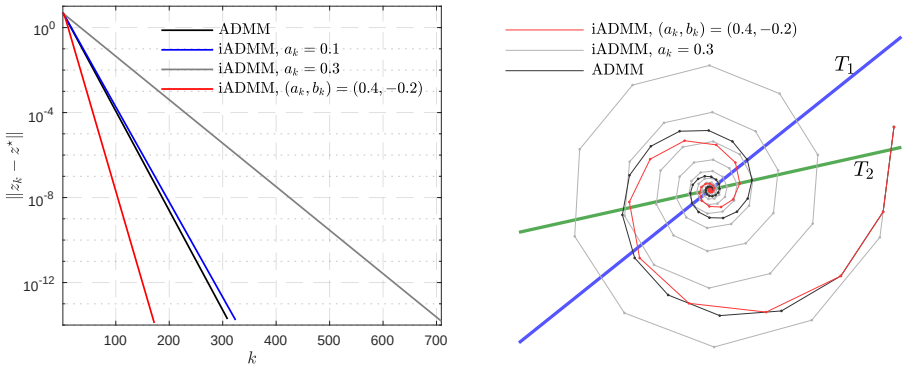
- Case $\gamma = \frac{\|K\|^2}{10}$: the value of $\cos(\theta_k)$ is plotted in Figure 2(a), from which we can observed that θ_k eventually is changing in an interval which implies that the leading eigenvalue of M_{ADMM} is complex, and the trajectory of z_k is a spiral. The inertial scheme works only 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 $\gamma = \frac{\|K\|^2}{10}$. As a result, the direction $z_k - z_{k-1}$ is not pointing towards $z^\star$, hence unable to provide satisfactory acceleration.
- Case $\gamma = \|K\|^2 + 0.1$: For this case, the leading eigenvalue of M_{ADMM} is real, hence the trajectory of z_k is straight line. From $k = 1000$, the increasing of $\cos(\theta_k)$ is due to machine errors. 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 $\gamma = \frac{\|K\|^2}{10}$ than $\gamma = \|K\|^2 + 0.1$, 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 $\gamma = \frac{\|K\|^2}{10}$ imply 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.

3.2 Feasibility problem

For the LASSO problem, though $a_k = 0.7, \frac{k-1}{k+2}$ fail to provide acceleration for $\gamma = \frac{\|K\|^2}{10}$, $a_k = 0.3$ is marginally faster than the standard ADMM. Below we consider a feasibility problem to demonstrate that (3.1) will fail to provide acceleration as long as $a_k > 0$.



(a) $\|z_k - z^\star\|$ (b) Trajectory of z_k

Figure 3: Performance of different schemes and trajectory of z_k for ADMM applying to feasibility problem.

Let T_1, T_2 be two subspaces of $\mathbb{R}^2$ such that $T_1 \cap T_2 = \{0\}$, and consider the following problem

$$\text{find } x \in \mathbb{R}^2 \quad \text{such that} \quad x \in T_1 \cap T_2, \tag{3.3}$$

which is finding the common point of T_1 and T_2. The problem can be written as

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} R(x) + J(y) \quad \text{such that} \quad x - y = 0,$$

where $R(x), J(y)$ are the indicator functions of T_1, T_2, respectively. Specify (3.1) to this case we obtain the

following iteration

$$\begin{aligned}
x_{k+1} &= \mathcal{P}_{T_1}\big((\bar{z}_k - 2\psi_k)/\gamma\big),\\
z_{k+1} &= \psi_k + \gamma x_{k+1},\\
\bar{z}_{k+1} &= z_{k+1} + a_k(z_{k+1} - z_k),\\
y_{k+1} &= \mathcal{P}_{T_2}(\bar{z}_{k+1}/\gamma),\\
\psi_{k+1} &= \bar{z}_{k+1} - \gamma y_{k+1}.
\end{aligned}$$

To demonstrate the failure of inertial ADMM, besides the standard ADMM, the following schemes are considered

$$\begin{aligned}
&\text{iADMM}: \bar{z}_k = z_k + 0.1(z_k - z_{k-1}),\\
&\text{iADMM}: \bar{z}_k = z_k + 0.3(z_k - z_{k-1}),\\
&\text{iADMM}: \bar{z}_k = z_k + 0.4(z_k - z_{k-1}) - 0.2(z_{k-1} - z_{k-2}).
\end{aligned}$$

The performance of the four schemes in terms of $\|z_k - z^\star\|$ is provided in Figure 3 (a). It can be observed that for the inertial schemes with $a_k = 0.1, 0.3$, they are both slower than the standard ADMM, while the last scheme is faster than all the others. Such a difference is due to the fact that the trajectory of $\{z_k\}_{k\in\mathbb{N}}$ of ADMM is a spiral, see Figure 3 (b) the black dot line. As a result, the direction $z_k - z_{k-1}$ is not pointing towards $z^\star$, making the inertial step useless. For the last inertial scheme, since the coefficient of $z_{k-1} - z_{k-2}$ is negative, the direction $0.4(z_k - z_{k-1}) - 0.2(z_{k-1} - z_{k-2})$ points towards $z^\star$, hence providing acceleration.

In Section A, substantial discussions on why inertial ADMM fails, when the trajectory of z_k is spiral, are provided. Especially for the case when both functions in ($\mathscr{P}_{\text{ADMM}}$) are (locally) polyhedral around the solution, *e.g.* the feasibility problem discussed above.

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

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

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 $V_{k-1} \overset{\text{def}}{=} \big[v_{k-1}, \cdots, v_{k-q}\big] \in \mathbb{R}^{n\times q}$, 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

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

which is the desired trajectory following extrapolation scheme. Define the extrapolation parameterized by s, q as

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

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 $\bar{q} = q+1$. Let $\bar{z}_0 = z_0 \in \mathbb{R}^p$ and $V_0 = 0 \in \mathbb{R}^{p\times q}$.

Repeat:

- For $k \geq 1$:

$$\begin{aligned} y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|By + \tfrac{1}{\gamma}(\bar{z}_{k-1} - \gamma b)\|^2, \\ \psi_k &= \bar{z}_{k-1} + \gamma(By_k - b), \\ x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax - \tfrac{1}{\gamma}(\bar{z}_{k-1} - 2\psi_k)\|^2, \\ z_k &= \psi_k + \gamma A x_k, \\ v_k &= z_k - z_{k-1} \quad \text{and} \quad V_k = [v_k, V_{k-1}(:, 1:q-1)]. \end{aligned} \tag{4.3}$$

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

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

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

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

Remark 4.1.

- When $\operatorname{mod}(k,\bar{q}) \neq 0$, one can also consider $\bar{z}_k = z_k + a_k(z_k - z_{k-1})$ with properly chosen a_k. In stead of $\bar{q} = q+1$, one can also consider $\bar{q} = q+i$ with $i \in \mathbb{N}_+$.
- A^{3}DMM carries out $\bar{q}$ standard ADMM iterations to set up the extrapolation step $\mathscr{E}_{s,q}$. As $\mathscr{E}_{s,q}$ contains the sum of the powers of C_k, it is guaranteed to be convergent when $\rho(C_k) < 1$. Therefore, we only apply $\mathscr{E}_{s,q}$ when the spectral radius $\rho(C_k) < 1$ is true. In this case, there is a closed form expression for $\mathscr{E}_{s,q}$ when $s = +\infty$; See Eq. (4.5).
- The purpose of adding a_k in front of $\mathscr{E}_{s,q}(z_k, \cdots, z_{k-q-1})$ is so that we can control the value of a_k to ensure the convergence of the algorithm; See below the discussion.

In Algorithm 1, we change the order of updates so that the extrapolation step only needs to be carried out on z_k. This is due to the fact, the update of y_k only depends on z_k, and such an arrangement requires the minimal computational overhead. Under such setting, the extra memory cost is pq for the storage of V_k. The extra computational cost of A^{3}DMM is very small, which is about nq^2 for computing the pseudo-inverse of V_k. Moreover, the value of q usually is taken very small, *e.g.* $q \leq 10$, therefore the total overheads of A^{3}DMM is rather limited.

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

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

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

$$z_k = \mathscr{F}(z_{k-1} + \varepsilon_{k-1}). \tag{4.4}$$

Owing to (4.4), 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 [6, Proposition 5.34].

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

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

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

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 $(\mathrm{Id} - C)^{-1} = \sum_{i=0}^{+\infty} C^i$. For the summation of the power of C_k in (4.2), when $s = +\infty$, we have

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

Back to (4.2), then we get

$$\begin{aligned} \bar{z}_{k,\infty} &\overset{\text{def}}{=} z_k + V_k\big((\mathrm{Id} - C_k)^{-1} - \mathrm{Id}\big)_{(:,1)} = z_k - v_k + V_k\big((\mathrm{Id} - C_k)^{-1}\big)_{(:,1)} \\ &= z_{k-1} + V_k\big((\mathrm{Id} - C_k)^{-1}\big)_{(:,1)} = \frac{1}{1 - \sum_{i=1}^s c_{k,i}}\big(z_k - \sum_{j=1}^{q-1} c_{k,j} z_{k-j}\big), \end{aligned} \tag{4.5}$$

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: $b \in \operatorname{argmin}_{a \in \mathbb{R}^{q+1}, \sum_j a_j = 1} \|\sum_{j=0}^q a_j v_{k-j}\|$ and $b_0 \neq 0$ and define $c_j \overset{\text{def}}{=} -b_j/b_0$ for $j = 1, \ldots, q$. Then,

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

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* $\varepsilon_k \overset{\text{def}}{=} \min_{c\in\mathbb{R}^q} \|V_{k-1}c - v_k\|$.

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

$$\|\bar{z}_{k,s} - z^\star\| \leq \|z_{k+s} - z^\star\| + B_s \varepsilon_k, \tag{4.6}$$

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 diagonalizable. 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 $k \to +\infty$): $\varepsilon_k = O(|\lambda_{q+1}|^k)$.*
- *Non-asymptotic 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,*

$$\frac{\varepsilon_k}{1 - \sum_i c_{k,i}} \leq K\beta^{k-q}\big(\tfrac{\sqrt{\eta}-1}{\sqrt{\eta}+1}\big)^q \tag{4.7}$$

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 (4.7) (see [49]), 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 $O(|\lambda_1|^k)$ to $O(|\lambda_{q+1}|^k)$, and the non-asymptotic bound shows that the improvement of extrapolation is optimal in the sense of Nesterov [41]. Recalling the form of the eigenvalues of M from Theorem 2.2, in the case of two non-smooth polyhedral terms, we must have $|\lambda_{2j-1}| = |\lambda_{2j}| > |\lambda_{2j+1}|$ for all $j \geq 1$. Hence, no acceleration can be guaranteed or observed when $q = 1$, while the choice of $q = 2$ provides guaranteed acceleration.

5 Discussions

In this section, two variants of ADMM, including the relaxed ADMM and symmetric ADMM which updates ψ twice every iteration, are discussed. An extension of A^{3}DMM Algorithm 1 to these variants is provided at the end of the section.

5.1 Variants of ADMM

Relaxed ADMM In the literature, a popular variant of ADMM is the *relaxed ADMM* which takes the following iteration procedure:

$$\begin{aligned} x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax + By_{k-1} - b + \tfrac{1}{\gamma}\psi_{k-1}\|^2, \\ \bar{x}_k &= \phi Ax_k - (1-\phi)(By_{k-1} - b), \\ y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|\bar{x}_k + By - b + \tfrac{1}{\gamma}\psi_{k-1}\|^2, \\ \psi_k &= \psi_{k-1} + \gamma(\bar{x}_k + By_k - b), \end{aligned} \tag{5.1}$$

where $\phi \in [0,2]$ is the relaxation parameter. The above iteration is called *over-relaxed* ADMM for $\phi \in]1,2]$.

In its dual form, the relaxed ADMM is equivalent to the *relaxed* Douglas–Rachford splitting applied to solve ($\mathscr{D}_{\text{ADMM}}$), see Section C.1.1. The convergence of (5.1) can be guaranteed for $\phi \in]0,2[$ [6]. Similar to (1.2), define $z_k \stackrel{\text{def}}{=} \psi_{k-1} + \gamma\bar{x}_k$, we can rewrite the relaxed ADMM into the following form

$$\begin{aligned} x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax - \tfrac{1}{\gamma}(z_{k-1} - 2\psi_{k-1})\|^2, \\ z_k &= \psi_{k-1} + \gamma\big(\phi Ax_k - (1-\phi)(By_{k-1} - b)\big), \\ y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|By + \tfrac{1}{\gamma}(z_k - \gamma b)\|^2, \\ \psi_k &= z_k + \gamma(By_k - b). \end{aligned} \tag{5.2}$$

We can easily adapt the result of the previous sections to the relaxed ADMM via z_k.

Remark 5.1. Similar to inertial acceleration, the performance of relaxation also depends on the trajectory of the sequence of z_k. For example, when both R, J are (locally) polyhedral around $x^\star, y^\star$, the (eventual) trajectory of z_k is spiral, according to [7] the (eventual) optimal relaxation parameter ϕ is 1, that is no relaxation provides the best performance.

Symmetric ADMM As aforementioned, the ADMM iteration (1.1) is equivalent to applying Douglas–Rachford splitting to the dual problem ($\mathscr{D}_{\text{ADMM}}$) [25]. It is also pointed out in [25] that, if the Peaceman–Rachford splitting method [42] is applied to solve ($\mathscr{D}_{\text{ADMM}}$), then it leads to the following iteration in the primal form

$$\begin{aligned} x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax + By_{k-1} - b + \tfrac{1}{\gamma}\psi_{k-1}\|^2, \\ \psi_{k-\frac{1}{2}} &= \psi_{k-1} + \gamma(Ax_k + By_{k-1} - b), \\ y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|Ax_k + By - b + \tfrac{1}{\gamma}\psi_{k-\frac{1}{2}}\|^2, \\ \psi_k &= \psi_{k-\frac{1}{2}} + \gamma(Ax_k + By_k - b), \end{aligned} \tag{5.3}$$

which is also called *symmetric ADMM*. A brief derivation is provided in Section C.1.2, and we refer to [25, 29] and the references therein for more detailed discussions.

In general, the conditions needed for the convergence of (5.3) is stronger than the standard ADMM (1.1), which is due to the fact that stronger conditions are needed to guarantee the convergence of Peaceman–Rachford splitting method [25]. However, when (5.3) converges, it tends to provide faster performance than (1.1). Similar to (1.2), if we define $z_k = \psi_k - \gamma By_k + \gamma b = \psi_{k-\frac{1}{2}} + \gamma Ax_k$, then (5.3) is equivalent to

$$\begin{aligned} x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax + \tfrac{1}{\gamma}(2\psi_{k-1} - z_{k-1})\|^2, \\ z_k &= \psi_{k-1} + \gamma(2Ax_k + By_{k-1} - b), \\ y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|By + \tfrac{1}{\gamma}(z_k - \gamma b)\|^2, \\ \psi_k &= z_k + \gamma(By_k - b), \end{aligned} \tag{5.4}$$

which can be written as the fixed-point iteration in terms of z_k, see Section C.1.2. Suppose the iteration is convergent, then following the analysis of Section 2, we can obtain the trajectory property of symmetric ADMM in terms of the fixed-point sequence z_k.

5.2 An extension of A^{3}DMM

From the above discussions, we have that relaxed ADMM (5.2) and symmetric ADMM (5.4) only differ from the standard ADMM on the update of z_k. As a result of these similarities, we can easily extend the A^{3}DMM Algorithm 1 to these variants.

Let

$$z_k = \mathscr{Z}(\gamma, \phi; x_k, y_{k-1}, \psi_{k-1})$$

represent the way of updating z_k in (1.2), (5.2) and (5.4). By replacing (4.3) in Algorithm 1 with the following equations

$$\begin{aligned} y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|By + \tfrac{1}{\gamma}(\bar{z}_{k-1} - \gamma b)\|^2, \\ \psi_k &= \bar{z}_{k-1} + \gamma(By_k - b), \\ x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax - \tfrac{1}{\gamma}(\bar{z}_{k-1} - 2\psi_k)\|^2, \\ z_k &= \mathscr{Z}(\gamma, \phi; x_k, y_{k-1}, \psi_{k-1}), \\ v_k &= z_k - z_{k-1} \quad \text{and} \quad V_k = [v_k, V_{k-1}(:, 1:q-1)], \end{aligned} \tag{5.5}$$

we obtain the extension of A^{3}DMM to the variants of ADMM.

6 Numerical experiments

We present numerical experiments on affine constrained minimisation (*e.g.* Basis Pursuit), LASSO, quadratic programming and image processing problems to demonstrate the performance of A^{3}DMM. 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: $(q,s) = (6,100)$ and $(q,s) = (6,+\infty)$. MATLAB source codes for reproducing the results can be found at: `https://github.com/jliang993/A3DMM`.

6.1 Affine constrained minimisation

Consider the following constrained problem

$$\min_{x\in\mathbb{R}^n} R(x) \quad \text{such that} \quad Kx = f. \tag{6.1}$$

Denote the set $\Omega \overset{\text{def}}{=} \{x \in \mathbb{R}^n : Kx = f\}$, and ι_Ω its indicator function. Then (6.1) can be written as

$$\min_{x,y\in\mathbb{R}^n} R(x) + \iota_\Omega(y) \quad \text{such that} \quad x - y = 0, \tag{6.2}$$

which is special case of ($\mathscr{P}_{\text{ADMM}}$) with $A = \text{Id}, B = -\text{Id}$ 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;

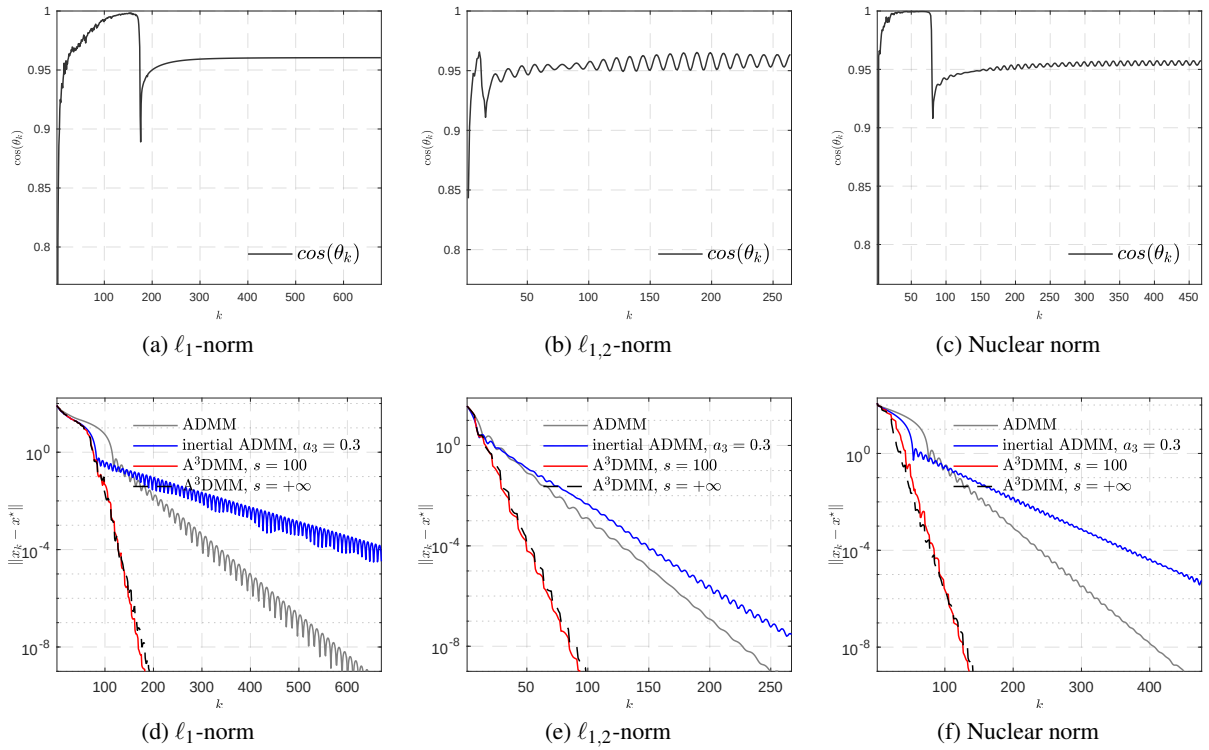
Nuclear norm $(m,n) = (1448, 4096)$, solution $x^\star$ has rank of 4.

The property of $\{\theta_k\}_{k\in\mathbb{N}}$ is shown in Figure 4 (a)-(c). Note that the indicator function $\iota_\Omega(y)$ in (6.2) is polyhedral since Ω is an affine subspace,

- As ℓ_1-norm is polyhedral, we have in Figure 4(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 4 (d)-(f):

- Since both functions in (6.2) 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

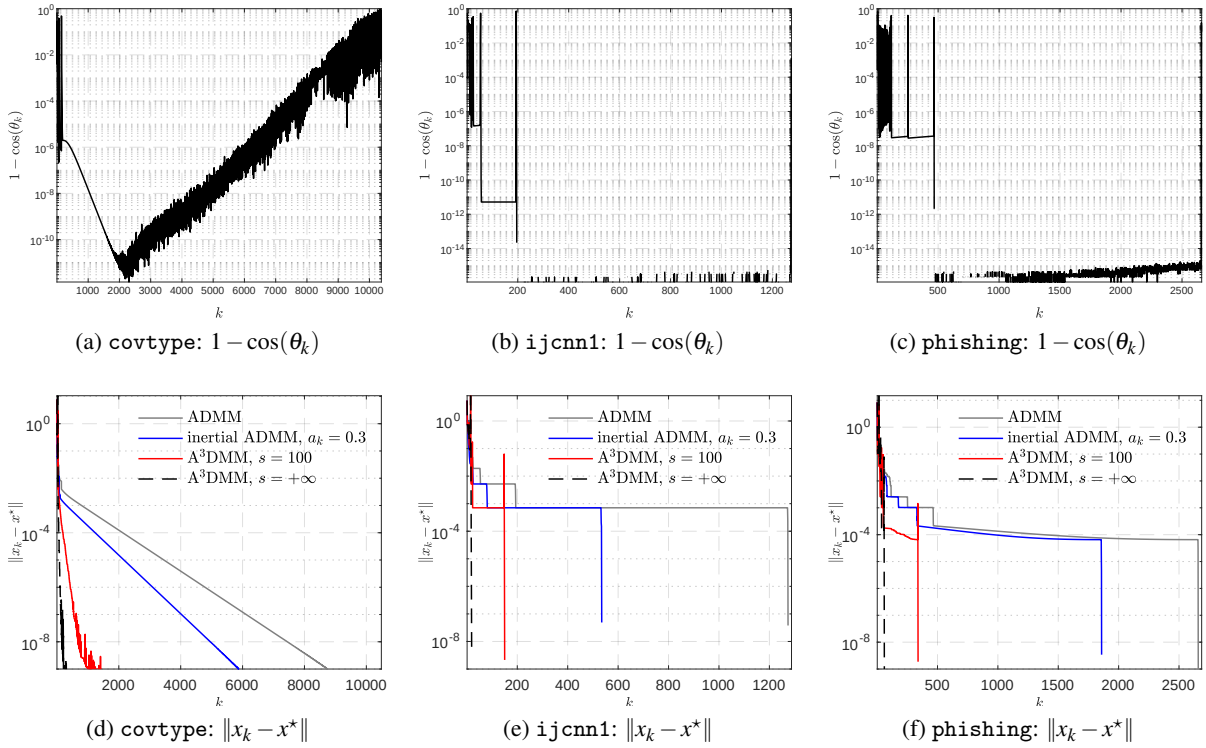
(d) ℓ_1-norm (e) $\ell_{1,2}$-norm (f) Nuclear norm

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

6.2 LASSO

We consider again the LASSO problem (3.2) with three datasets from LIBSVM[1]. The numerical experiments are provided below in Figure 5.

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.



(a) `covtype`: $1-\cos(\theta_k)$ (b) `ijcnn1`: $1-\cos(\theta_k)$ (c) `phishing`: $1-\cos(\theta_k)$

(d) `covtype`: $\|x_k - x^\star\|$ (e) `ijcnn1`: $\|x_k - x^\star\|$ (f) `phishing`: $\|x_k - x^\star\|$

Figure 5: Performance comparisons for LASSO problem.

6.3 Quadratic programming

Consider the following quadratic optimisation problem

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \frac{1}{2}x^T Qx + \langle q, x\rangle, \\ \text{such that} \quad & x_i \in [\ell_i, r_i],\ i = 1,\ldots,n. \end{aligned} \tag{6.3}$$

Define the constraint set $\Omega = \{x \in \mathbb{R}^n : x_i \in [\ell_i, r_i],\ i = 1,\ldots,n\}$, then (6.3) can be written as

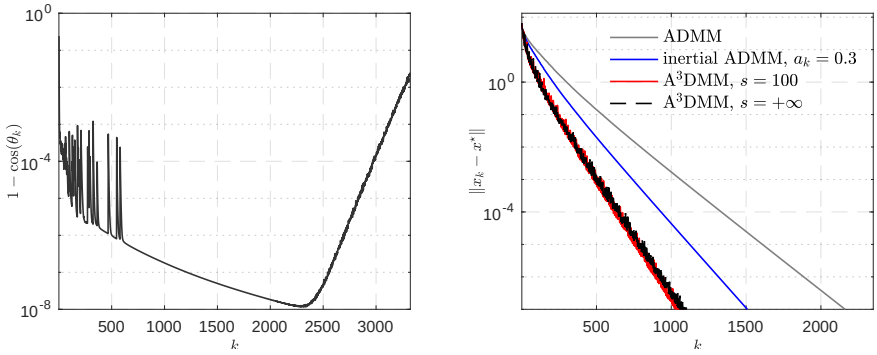
$$\min_{x,y\in\mathbb{R}^n} \quad \frac{1}{2}x^T Qx + \langle q, x\rangle + \iota_\Omega(y) \quad \text{such that} \quad x - y = 0,$$

which is special case of ($\mathscr{P}_{\text{ADMM}}$) with $A = \text{Id}, B = -\text{Id}$ and $b = 0$.

The angle θ_k of ADMM and the performances of the four schemes are provided in Figure (6), 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 previous observations, the proposed A^{3}DMM schemes provides the best performance.

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



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

Figure 6: Performance comparisons and $\{\theta_k\}_{k\in\mathbb{N}}$ of ADMM for quadratic programming.

6.4 Total variation based image inpainting

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

$$\min_{x\in\mathbb{R}^{n\times n}} \|\nabla x\|_1 \quad \text{such that} \quad \mathcal{P}_{\mathscr{S}}(x) = f. \tag{6.4}$$

Define $\Omega \overset{\text{def}}{=} \{x \in \mathbb{R}^{n\times n} : \mathcal{P}_{\mathscr{S}}(x) = f\}$, then (6.4) becomes

$$\min_{x\in\mathbb{R}^{n\times n}, y\in\mathbb{R}^{2n\times n}} \|y\|_1 + \iota_\Omega(x) \quad \text{such that} \quad \nabla x - y = 0, \tag{6.5}$$

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

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

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

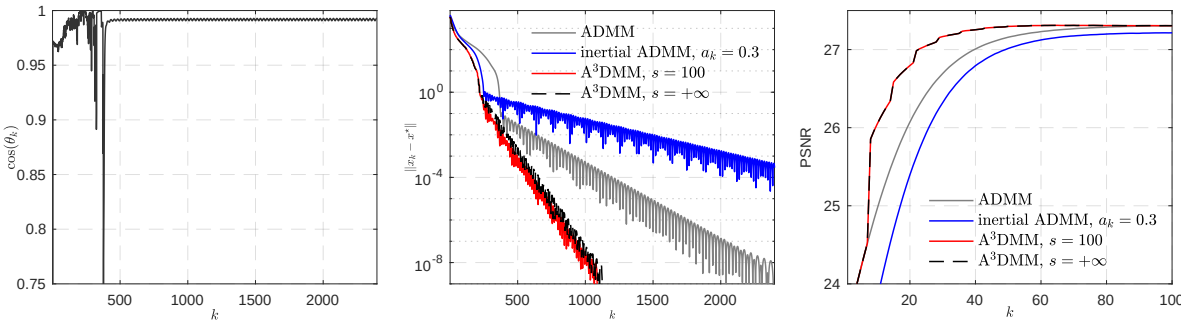
In the experiment, the `cameraman` image is used, and 50% of the pixels is removed randomly. The angle θ_k of ADMM and the comparisons of the four schemes are provided in Figure 7:

- Though both functions in (6.5) 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 7(c) implies that A^{3}DMM is also the best.

We also compare the visual quality of the images obtained by the four schemes for the 30'th iteration, which is shown below in Figure 8. It can be observed that the image quality (2nd row of Figure 8) is much better than the 1st row of ADMM and inertial ADMM.

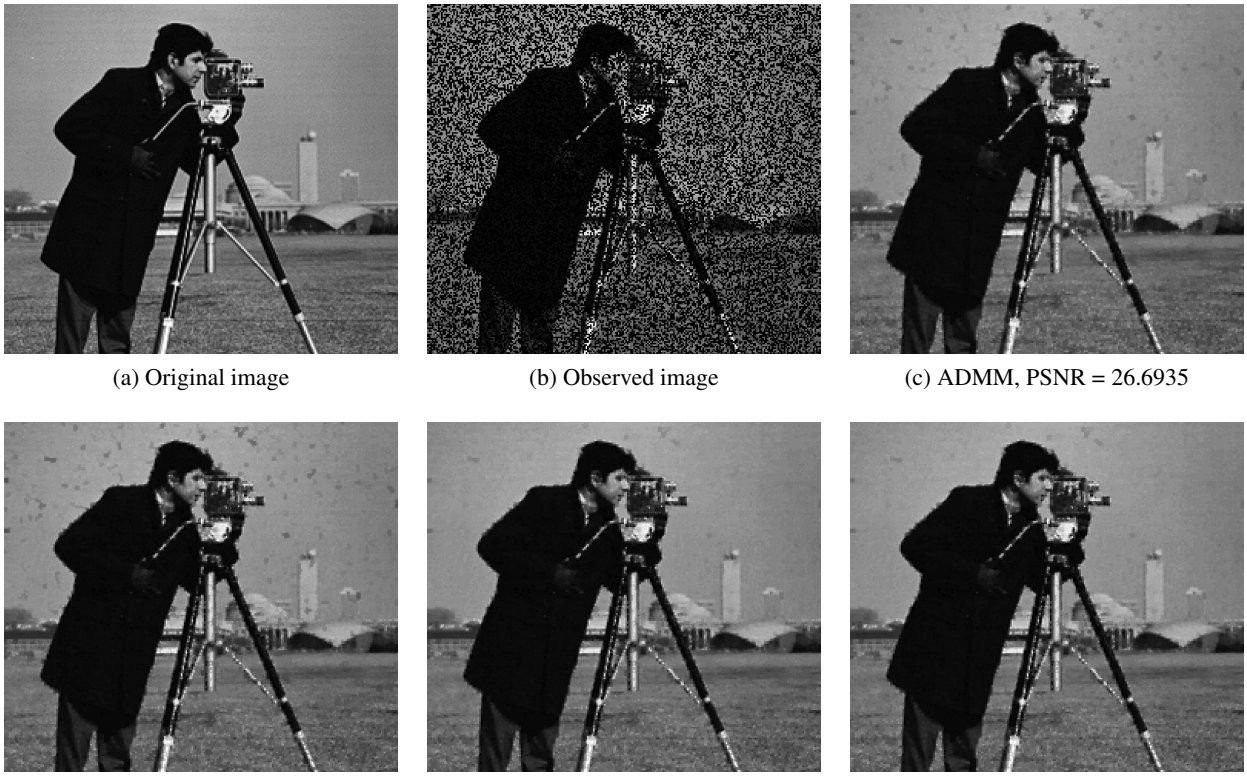
7 Conclusions

In this article, by analyzing 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 framework to general first-order methods is ongoing.



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

Figure 7: Property of $\{\theta_k\}_{k\in\mathbb{N}}$, performance comparison and image quality of ADMM for TV based image inpainting.



(a) Original image (b) Observed image (c) ADMM, PSNR = 26.6935

(d) Inertial ADMM, PSNR = 26.3203 (e) A^{3}DMM $s = 100$, PSNR = 27.1668 (f) A^{3}DMM $s = +\infty$, PSNR = 27.1667

Figure 8: Comparison of image quality at the 30'th iteration of ADMM, inertial ADMM and A^{3}DMM with two different prediction steps.

Acknowledgements

We would like to thank Arieh Iserles for pointing out the connection between trajectory following adaptive acceleration and vector extrapolation. We also like to thank the reviewers whose comments helped to improve the paper. Jingwei Liang was partly supported by Leverhulme trust, Newton trust, the EPSRC centre "EP/N014588/1" and the Cantab Capital Institute for the Mathematics of Information (CCIMI).

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] B. H. Bauschke and D. Noll. On the local convergence of the douglas–rachford algorithm. *Archiv der Mathematik*, 102(6):589–600, 2014.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

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

[14] 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.

[15] 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.

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

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

[18] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

[19] 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.

[20] J. Eckstein and D. P. Bertsekas. On the douglas–rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[21] J. Eckstein and W. Yao. Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results. *RUTCOR Research Reports*, 32(3), 2012.

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

[23] G. Franca, D. P. Robinson, and R. Vidal. A dynamical systems perspective on nonsmooth constrained optimization. *arXiv preprint arXiv:1808.04048*, 2018.

[24] Guilherme Franca, Daniel Robinson, and Rene Vidal. ADMM and accelerated ADMM as continuous dynamical systems. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1559–1567, Stockholmsmassan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

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

[26] D. Gabay and B. Mercier. *A dual algorithm for the solution of non linear variational problems via finite element approximation*. Institut de recherche d'informatique et d'automatique, 1975.

[27] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

[28] 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.

[29] B. He, H. Liu, Z. Wang, and X. Yuan. A strictly contractive peaceman–rachford splitting method for convex programming. *SIAM Journal on Optimization*, 24(3):1011–1040, 2014.

[30] B. He and X. Yuan. On the $o(1/n)$ convergence rate of the douglas–rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[31] M. Hong and Z. Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.

[32] 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.

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

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

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] D. W. Peaceman and H. H. Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for Industrial & Applied Mathematics*, 3(1):28–41, 1955.

[43] 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.

[44] B. T. Polyak. *Introduction to optimization*. Optimization Software, 1987.

[45] 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.

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

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

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

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

[50] 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.

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

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

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

Appendix

The organization of the appendix is as follows: In Section A we provide more discussions on the conditions when inertial fails. The proofs of the main results of the paper are contained in Sections B-4, where in Section B some preliminary result on angles between subspaces and Riemannian geometry are provided, in Section C the proofs for the trajectory of ADMM are provided, and lastly in in Section D we provide proofs for A^{3}DMM.

A The failure of inertial acceleration continue

In this part, to support the discussion of Section 3, we provide extra discussion on why inertial acceleration, in particular Nesterov/FISTA, will fail when the (leading) eigenvalue of M is complex.

Let $M \in \mathbb{R}^{n\times n}$ be a square matrix and consider the following linear equation

$$z_{k+1} = Mz_k. \tag{A.1}$$

According to [44], (A.1) is linearly convergent when the spectral radius of M is strictly smaller than 1, *i.e.* $\rho(M) < 1$. For simplicity, consider the inertial version of (A.1) with fixed inertial parameter $a_k \equiv a \in [0,1]$, we get

$$\begin{aligned} y_k &= z_k + a(z_k - z_{k-1}) \\ z_{k+1} &= My_k. \end{aligned} \tag{A.2}$$

The above scheme corresponds to the local linearization of the inertial ADMM (3.1) without the small o-term. Define the augmented variable $w_k = \begin{pmatrix} z_k \\ z_{k-1} \end{pmatrix}$ and block matrix $\widetilde{M} \overset{\text{def}}{=} \begin{bmatrix} (1+a)M & -aM \\ \mathrm{Id} & 0 \end{bmatrix}$, then (A.2) can be written as

$$w_{k+1} = \widetilde{M} w_k. \tag{A.3}$$

To guarantee the convergence of (A.3), we require the spectral radius satisfying $\rho(\widetilde{M}) < 1$. Therefore, in the following, motivated by [44, 35, 37], we discuss the property of the spectral radius $\rho(\widetilde{M})$ and the conditions such that $\rho(\widetilde{M}) < 1$.

Let η, ρ be the leading eigenvalues of M and $\widetilde{M}$, respectively. According to [37, Proposition 4.6], we have the following lemma regarding the relation between η and ρ.

Lemma A.1 ([37, Proposition 4.6]). *Suppose* $\begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$ *is the eigenvector of* $\widetilde{M}$ *corresponding to eigenvalue* ρ*, then it must satisfy* $r_1 = \rho r_2$*. Moreover,* r_2 *is an eigenvector of* M *associated to eigenvalue* η*, where* η *and* ρ *satisfy the relation*

$$\rho^2 - (1+a)\eta\rho + a\eta = 0. \tag{A.4}$$

The relation (A.4) is a simple quadratic equation of ρ, we have

$$\rho = \frac{(1+a)\eta + \sqrt{(1+a)^2\eta^2 - 4a\eta}}{2}. \tag{A.5}$$

The value of $|\rho|$ depends on a and η, and the discussion splits into two scenarios: η is real and η is complex.

A.1 Real η

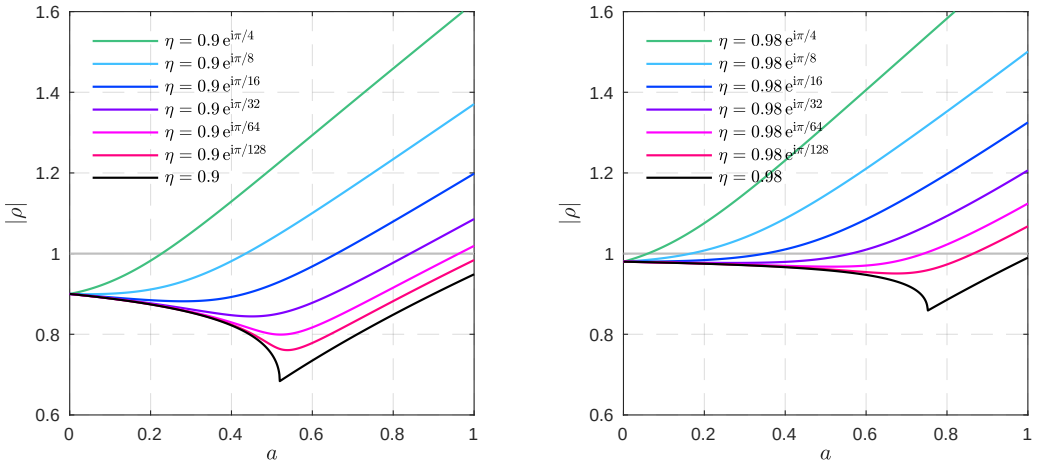
When η is real valued, the property of ρ is well studied, we refer to [37] and references therein for detailed discussions. Basically, we have that

$$|\rho| = \begin{cases} (1+a)^2\eta^2 \geq 4a\eta : \rho \text{ is real, } |\rho| < 1 \text{ holds for any } a \in [0,1], \\ (1+a)^2\eta^2 < 4a\eta : \rho \text{ is complex, } |\rho| = \sqrt{a\eta} < 1 \text{ holds for any } a \in [0,1]. \end{cases}$$

The above result can be summarized below.

Lemma A.2 ([37, Proposition 4.6]). *Given any $a \in [0,1]$, we have $|\rho| < 1$ as long as $0 \leq \eta < 1$.*

To demonstrate the above result, we consider fixing η and varying $a \in [0,1]$. Two choices of η are considered $\eta = 0.9, 0.98$, the value of $|\rho|$ is plotted in Figure 9 in black line. It can be observed that $|\rho|$ is strictly smaller than one for both choices of η. Note that $|\rho|$ reaches a minimal value for some a, we refer to [37] for detailed discussion on this.



(a) $|\eta| = 0.9$ (b) $|\eta| = 0.98$

Figure 9: The value of $|\rho|$ under fixed $|\eta|$ and $a \in [0,1]$.

A.2 Complex η

When η is complex, it can be written as $\eta = |\eta|e^{i\alpha}$ where α is the argument of η. The dependence of $|\rho|$ on a and η becomes much more complicated, below we briefly demonstrate where the difficulties arise and provide numerical proof for the properties of $|\rho|$.

General form $\eta = |\eta|e^{i\alpha}$ For this case, we have

$$\rho = \frac{(1+a)\eta + \sqrt{(1+a)^2\eta^2 - 4a\eta}}{2} = \frac{(1+a)|\eta|e^{i\alpha} + \sqrt{(1+a)^2|\eta|^2e^{i2\alpha} - 4a|\eta|e^{i\alpha}}}{2}.$$

Suppose $(x+iy)^2 = (1+a)^2|\eta|^2e^{i2\alpha} - 4a|\eta|e^{i\alpha}$, we get

$$x^2 - y^2 = (1+a)^2|\eta|^2\cos(2\alpha) - 4a|\eta|\cos(\alpha)$$
$$xy = \frac{(1+a)^2|\eta|^2\sin(2\alpha) - 4a|\eta|\sin(\alpha)}{2},$$

which can be simplified to a equation of x

$$x^4 - \big((1+a)^2|\eta|^2\cos(2\alpha) - 4a|\eta|\cos(\alpha)\big)x^2 - \frac{((1+a)^2|\eta|^2\sin(2\alpha) - 4a|\eta|\sin(\alpha))^2}{4} = 0.$$

Solving the above equation, we get

$$x = \left(\frac{((1+a)^2|\eta|^2\cos(2\alpha) - 4a|\eta|\cos(\alpha)) + \sqrt{((1+a)^2|\eta|^2\cos(2\alpha) - 4a|\eta|\cos(\alpha))^2 + ((1+a)^2|\eta|^2\sin(2\alpha) - 4a|\eta|\sin(\alpha))^2}}{2}\right)^{1/2},$$
$$y = \frac{(1+a)^2|\eta|^2\sin(2\alpha) - 4a|\eta|\sin(\alpha)}{2x},$$

here we only take the positive root x. Back to the expression of ρ, we get

$$\rho = \frac{(1+a)|\eta|e^{i\alpha} + (x+iy)}{2} = \frac{((1+a)|\eta|\cos(\alpha) + x) + i((1+a)|\eta|\sin(\alpha) + y)}{2}.$$

Given the complicated form of x, the analysis of $|\rho|$ becomes rather difficult. Therefore, below we discuss the property of $|\rho|$ through numerical verification.

Similar to the real η case, $|\eta| = 0.9, 0.98$ are considered. Denote α the argument of η, then we have $\eta = |\eta|e^{i\alpha}$. In total, six choices of α are considered: $\alpha \in \{\frac{\pi}{4}, \frac{\pi}{8}, \frac{\pi}{16}, \frac{\pi}{32}, \frac{\pi}{64}, \frac{\pi}{128}\}$. The value of $|\rho|$ are shown in Figure 9. Taking Figure 9 (a) for example, we have the following observations:

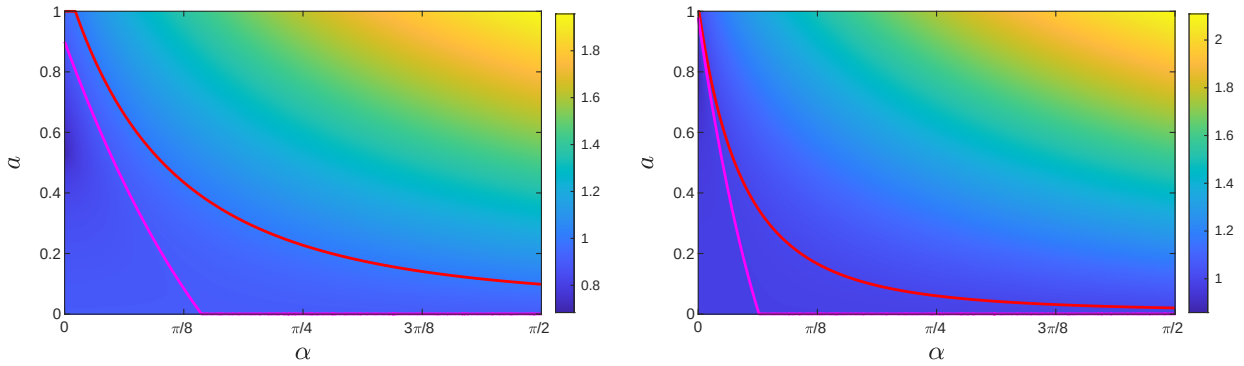
- For all choices of α except $\alpha = \frac{\pi}{128}$, there exists an $a_\alpha < 1$ such that $|\rho| \geq 1$ for $a \in [a_\alpha, 1]$.
- The larger the value of α, the smaller the value of a_α, see the green line in both figures.

From the above discussion, we can conclude that

- The inertial scheme is robust when all the eigenvalues of M are real, and we can afford the inertial parameter up to 1 which includes the FISTA [9] schemes as $a_k \to 1$, same for the Nesterov's accelerated gradient descent.
- When M has complex eigenvalue(s), which is not necessary to the leading eigenvalue, the largest value of a can be allowed is smaller than 1 and FISTA/Nesterov's scheme will fail.

To complete the discussion, we consider the values of $|\rho|$ under $\alpha \in [0, \pi/2]$ and $a \in [0, 1]$. The results are shown below in Figure 10. Again $|\eta| = 0.9, 0.98$ are considered. The horizontal axis is for α while the vertical is for a, each point inside the square stands for the value of $|\rho|$ with colorbar provided. In each figure:

- The *red* line stands for $|\rho| = 1$. Therefore, only for the area below the red line we have $|\rho| < 1$. Given any $\alpha \in [0, \pi/2]$, the larger the value of α, the smaller range of choice of a such that $|\rho| < 1$. This coincides with the observations from Figure 9.
- The *magenta* line stands for $|\rho| = |\eta|$. Only the small area below the magenta line has $|\rho| < |\eta|$, meaning that acceleration can be obtained. As a result, given $\eta = |\eta|e^{i\alpha}$, when α is large enough, such as about $\pi/8$ for $|\eta| = 0.9$, inertial will fail to provide acceleration.

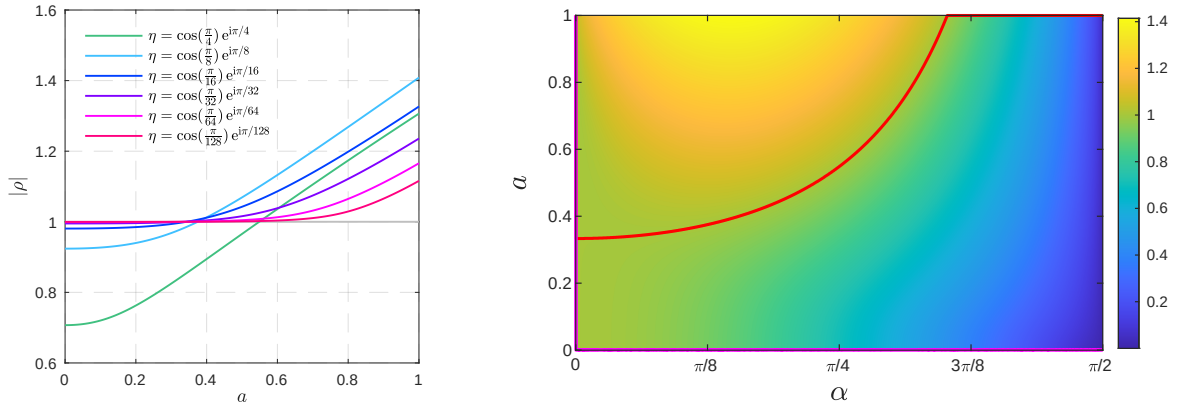


(a) $|\eta| = 0.9$ (b) $|\eta| = 0.98$

Figure 10: The value of $|\rho|$ under fixed η and $a \in [0, 1]$.

It should be noted that, for the above discussion, we consider the case that the leading eigenvalue is *complex*, while the rest of the eigenvalues are *real*. For the case leading eigenvalue is *real* while the rest are *complex*, then the spectral

radius of $\widetilde{M}$ will be determined by the non-leading complex eigenvalues when the inertial parameter a is large enough. Consequently, the FISTA inertial parameter rule still can not be applied, unless the magnitude of the leading eigenvalue is small enough; See Figure 10 (a).



(a) $\alpha \in \{\frac{\pi}{4}, \frac{\pi}{8}, \frac{\pi}{16}, \frac{\pi}{32}, \frac{\pi}{64}, \frac{\pi}{128}\}$ (b) $\alpha \in [0, \frac{\pi}{2}]$

Figure 11: The value of $|\rho|$ when $\eta = \cos(\alpha)e^{i\alpha}$ and $a \in [0,1]$.

Special case $\eta = \cos\alpha e^{i\alpha}$ Now we consider a special case where $\eta = \cos(\alpha)e^{i\alpha}, \alpha \in [0,\pi/2]$ which corresponds to the case R,J in ($\mathscr{P}_{\text{ADMM}}$) are locally polyhedral around $x^\star, y^\star$. Similar to above, six choices of α are considered: $\alpha \in \{\frac{\pi}{4}, \frac{\pi}{8}, \frac{\pi}{16}, \frac{\pi}{32}, \frac{\pi}{64}, \frac{\pi}{128}\}$. The value of $|\rho|$ is shown below in Figure 11 (a). It can be observed that, for each α, the value of $|\rho|$ is monotonically increasing as the value of a increases, which means *inertial slows down the speed of convergence*. In Figure 11 (b), we consider the value of $|\rho|$ under $\alpha \in [0,\pi/2]$ and $a \in [0,1]$. We have

- Similar to Figure 10, the *red* line stands for $|\rho| = 1$. For each α, $|\rho| < 1$ for all the choices of a under the red line.
- The *magenta* line stands for $|\rho| = |\eta|$. It can be observed that, except for $\alpha = 0$ where $|\rho| = 1$ holds for all $a \in [0,1]$, $|\rho| = 1$ holds only for $a = 0$ when $\alpha \in]0,\pi/2]$.

Therefore, we can conclude that when R,J are locally polyhedral around the solution $x^\star, y^\star$, inertial scheme will not provide any acceleration.

B Preparatory materials

B.1 Polynomial extrapolation

Minimal polynomial extrapolation (MPE) [14]: 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) [22, 39] is obtained by replacing the first step by

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

The motivation for the use of such methods for the acceleration of fixed point sequences $x_{k+1} = \mathscr{F}(z_k)$ come from considering the spectral properties of the linearization 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 [50, 51, 49] for details on these methods and their acceleration guarantees.

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

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

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

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

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

The following lemma shows the relation between the Friedrichs and principal angles, whose proof can be found in [7, Proposition 3.3].

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

B.3 Riemannian Geometry

Let $\mathscr{M}$ 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 $\mathscr{M}$ into $\mathbb{R}^n$ permits to define a Riemannian structure and to introduce geodesics on $\mathscr{M}$, and we simply say $\mathscr{M}$ is a Riemannian manifold. We denote respectively $\mathscr{T}_{\mathscr{M}}(x)$ and $\mathscr{N}_{\mathscr{M}}(x)$ the tangent and normal space of $\mathscr{M}$ at point near x in $\mathscr{M}$.

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 $\mathscr{M}$. We denote by $\mathfrak{g}(t; x, h)$ the value at $t \in \mathbb{R}$ of the geodesic starting at $\mathfrak{g}(0; x, h) = x \in \mathscr{M}$ with velocity $\dot{\mathfrak{g}}(t; x, h) = \frac{d\mathfrak{g}}{dt}(t; x, h) = h \in \mathscr{T}_{\mathscr{M}}(x)$ (which is uniquely defined). For every $h \in \mathscr{T}_{\mathscr{M}}(x)$, there exists an interval I around 0 and a unique geodesic $\mathfrak{g}(t; x, h) : I \to \mathscr{M}$ such that $\mathfrak{g}(0; x, h) = x$ and $\dot{\mathfrak{g}}(0; x, h) = h$. The mapping

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

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

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

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

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

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 $v \in \mathscr{N}_{\mathscr{M}}(x)$, the Weingarten map of $\mathscr{M}$ at x is the operator $\mathfrak{W}_x(\cdot, v) : \mathscr{T}_{\mathscr{M}}(x) \to \mathscr{T}_{\mathscr{M}}(x)$ defined by

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

where V is any local extension of v to a normal vector field on $\mathscr{M}$. 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 $\mathscr{M}$, see [16, Proposition II.2.1].

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

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

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

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

This definition agrees with the usual definition using geodesics or connections [40]. Now assume that $\mathscr{M}$ is a Riemannian embedded submanifold of $\mathbb{R}^n$, and that a function G has a C^2-smooth restriction on $\mathscr{M}$. 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 $\mathscr{M}$. Thus, the Riemannian gradient $\nabla_{\mathscr{M}} G(x')$ is also given by

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

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

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

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

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

See [33, 16] for more materials on differential and Riemannian manifolds.

B.4 Preparatory lemmas

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

Lemma B.4 ([36, Lemma 5.1]). *Let $\mathscr{M}$ be a C^2-smooth manifold around x. Then for any $x' \in \mathscr{M} \cap \mathscr{N}$, where $\mathscr{N}$ is a neighborhood of x, the projection operator $\mathcal{P}_{\mathscr{M}}(x')$ is uniquely valued and C^1 around x, and thus*

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

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

Lemma B.5 ([37, Lemma B.1]). *Let $x \in \mathscr{M}$, and x_k a sequence converging to x in $\mathscr{M}$. Denote $\tau_k : \mathscr{T}_{\mathscr{M}}(x_k) \to \mathscr{T}_{\mathscr{M}}(x)$ 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*

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

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

Lemma B.6 ([37, Lemma B.2]). *Let x, x' be two close points in $\mathscr{M}$, denote $\tau : \mathscr{T}_{\mathscr{M}}(x') \to \mathscr{T}_{\mathscr{M}}(x)$ the parallel translation along the unique geodesic joining x to x'. The Riemannian Taylor expansion of $R \in C^2(\mathscr{M})$ around x reads,*

$$\tau \nabla_{\mathscr{M}} R(x') = \nabla_{\mathscr{M}} R(x) + \nabla^2_{\mathscr{M}} R(x) \mathcal{P}_{\mathscr{T}_{\mathscr{M}}(x)}(x' - x) + o(\|x' - x\|). \tag{B.4}$$

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

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

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

$$\nabla^2_{\mathscr{M}_x} R(x')h = \mathcal{P}_{T_{x'}} \nabla^2 \widetilde{R}(x')h + \mathfrak{W}_{x'}\big(h, \mathcal{P}_{T_{x'}^{\perp}} \nabla \widetilde{R}(x')\big),$$

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

B.5 Linearization of proximal mapping

In this part, we present one fundamental result led by partial smoothness, the linearization 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 $\mathscr{M}_{\bar{x}}$ and $\bar{u} \in \partial R(\bar{x})$, define the following smooth perturbation of R

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

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

Lemma B.8 ([37, Lemma 4.2]). *Let $R \in \Gamma_0(\mathbb{R}^n)$ be partly smooth at $\bar{x}$ relative to $\mathscr{M}_{\bar{x}}$, then $H_{\bar{R}}$ is symmetric positive semi-definite if either of the following is true:*

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

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

One consequence of Lemma B.8 is that, we can linearize the generalized 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 generalized 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. Denote $A_{T_{\bar{x}}} \stackrel{\text{def}}{=} A \circ \mathcal{P}_{T_{\bar{x}}}$, it is immediate that $A^T_{T_{\bar{x}}} A_{T_{\bar{x}}}$ is positive semidefinite and invertible along $T_{\bar{x}}$. In the following we denote $(A^T_{T_{\bar{x}}} A_{T_{\bar{x}}})^{-1}$ the inverse along $T_{\bar{x}}$ Denote

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

Lemma B.9. *Let function $R \in \Gamma_0(\mathbb{R}^n)$ be partly smooth at the point $\bar{x}$ relative to the manifold $\mathscr{M}_{\bar{x}}$ 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 hold $x_k \in \mathscr{M}_{\bar{x}}$ and*

$$A_{T_{\bar{x}}}(x_k - x_{k-1}) = M_{\bar{R}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|). \tag{B.5}$$

Remark B.10. When $A = \mathrm{Id}$, then $\mathrm{prox}^A_{\gamma R}$ reduces to the standard proximal mapping, and (B.5) simplifies to

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

In [35] and references therein, to study the local linear convergence of first-order methods, linearization with respect to the limiting points is provided, that is

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

Proof. Since R is proper convex and lower semi-continuous, we have $R(x_k) \to R(\bar{x})$ and $\partial R(x_k) \ni u_k = -A^T(Ax_k - w_k)/\gamma \to \bar{u} \in \mathrm{ri}(\partial R(\bar{x}))$, hence $\mathrm{dist}(u_k, \partial R(\bar{x})) \to 0$. As a result, we have $x_k \in \mathscr{M}_{\bar{x}}$ owing to [28, Theorem 5.3] and $u_k \in \mathrm{ri}(\partial R(x_k))$ owing to [52] for all k large enough.

Denote $T_{x_k}, T_{x_{k-1}}$ the tangent spaces of $\mathscr{M}_{\bar{x}}$ 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 $\mathscr{M}_{\bar{x}}$ joining x_k to x_{k-1}. From the definition of x_k, let $h_k = \gamma u_k$, we get

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

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

$$\begin{aligned}
&\gamma\tau_k \nabla_{\mathscr{M}_{\bar{x}}} R(x_k) = \tau_k \mathcal{P}_{T_{x_k}}(h_k) = \mathcal{P}_{T_{x_{k-1}}}(h_k) + \big(\tau_k \mathcal{P}_{T_{x_k}} - \mathcal{P}_{T_{x_{k-1}}}\big)(h_k),\\
&\gamma\nabla_{\mathscr{M}_{\bar{x}}} R(x_{k-1}) = \mathcal{P}_{T_{x_{k-1}}}(h_{k-1}).
\end{aligned}$$

The difference of the above two equalities yields

$$\gamma\tau_k \nabla_{\mathscr{M}_{\bar{x}}} R(x_k) - \gamma\nabla_{\mathscr{M}_{\bar{x}}} R(x_{k-1}) - \big(\tau_k \mathcal{P}_{T_{x_k}} - \mathcal{P}_{T_{x_{k-1}}}\big)(h_{k-1}) = \mathcal{P}_{T_{x_{k-1}}}(h_k - h_{k-1}) + \big(\tau_k \mathcal{P}_{T_{x_k}} - \mathcal{P}_{T_{x_{k-1}}}\big)(h_k - h_{k-1}). \tag{B.6}$$

Owing to the monotonicity of subdifferential, *i.e.* $\langle h_k - h_{k-1}, x_k - x_{k-1}\rangle \geq 0$, we get

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

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

$$\|h_k - h_{k-1}\| = \|A^T(Ax_k - w_k) - A^T(Ax_{k-1} - w_{k-1})\| \leq \|A\|^2\|x_k - x_{k-1}\| + \|A\|\|w_k - w_{k-1}\| \leq \big(\tfrac{\|A\|^3}{\kappa} + \|A\|\big)\|w_k - w_{k-1}\|.$$

As a result, owing to Lemma B.5, we have for the term $(\tau_k \mathcal{P}_{T_{x_k}} - \mathcal{P}_{T_{x_{k-1}}})(h_k - h_{k-1})$ in (B.6) that

$$\big(\tau_k \mathcal{P}_{T_{x_k}} - \mathcal{P}_{T_{x_{k-1}}}\big)(h_k - h_{k-1}) = o(\|h_k - h_{k-1}\|) = o(\|w_k - w_{k-1}\|).$$

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

$$\begin{aligned}
\gamma\tau_k \nabla_{\mathscr{M}_{\bar{x}}} R(x_k) - \gamma \nabla_{\mathscr{M}_{\bar{x}}} R(x_{k-1}) - \big(\tau_k \mathcal{P}_{T_{x_k}} - \mathcal{P}_{T_{x_{k-1}}}\big)(h_{k-1}) &= \tau_k\big(\gamma \nabla_{\mathscr{M}_{\bar{x}}} R(x_k) - \mathcal{P}_{T_{x_k}}(h_{k-1})\big) - \big(\gamma \nabla_{\mathscr{M}_{\bar{x}}} R(x_{k-1}) - \mathcal{P}_{T_{x_{k-1}}}(h_{k-1})\big) \\
&= \tau_k \nabla_{\mathscr{M}_{\bar{x}}} \bar{R}_{k-1}(x_k) - \nabla_{\mathscr{M}_{\bar{x}}} \bar{R}_{k-1}(x_{k-1}) \\
&= H_{\bar{R},k-1}(x_k - x_{k-1}) + o(\|x_k - x_{k-1}\|) \\
&= H_{\bar{R},k-1}(x_k - x_{k-1}) + o(\|w_k - w_{k-1}\|).
\end{aligned} \tag{B.7}$$

Back to (B.6), we get

$$H_{\bar{R},k-1}(x_k - x_{k-1}) = \mathcal{P}_{T_{x_{k-1}}}(h_k - h_{k-1}) + o(\|w_k - w_{k-1}\|). \tag{B.8}$$

Define $\bar{R}(x) \overset{\text{def}}{=} \gamma R(x) - \langle x, \bar{h}\rangle$ and $H_{\bar{R}} = \mathcal{P}_{T_{\bar{x}}} \nabla^2_{\mathscr{M}_{\bar{x}}} \bar{R}(\bar{x}) \mathcal{P}_{T_{\bar{x}}}$, then from (B.8) that

$$H_{\bar{R}}(x_k - x_{k-1}) + \big(H_{\bar{R},k-1} - H_{\bar{R}}\big)(x_k - x_{k-1}) = \mathcal{P}_{T_{\bar{x}}}(h_k - h_{k-1}) + \big(\mathcal{P}_{T_{x_{k-1}}} - \mathcal{P}_{T_{\bar{x}}}\big)(h_k - h_{k-1}) + o(\|w_k - w_{k-1}\|). \tag{B.9}$$

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

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

and $\lim_{k\to+\infty} \frac{\|(\mathcal{P}_{T_{x_{k-1}}} - \mathcal{P}_{T_{\bar{x}}})(x_k - x_{k-1})\|}{\|x_k - x_{k-1}\|} = 0$. 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 B.4, and denoting $A_{T_{\bar{x}}} = A \circ \mathcal{P}_{T_{\bar{x}}}$, equation (B.9) can be written as

$$\begin{aligned}
H_{\bar{R}}(x_k - x_{k-1}) = \mathcal{P}_{T_{\bar{x}}}(u_k - u_{k-1}) + o(\|w_k - w_{k-1}\|) &= -\mathcal{P}_{T_{\bar{x}}}(A^T(Ax_k - w_k) - A^T(Ax_{k-1} - w_{k-1})) + o(\|w_k - w_{k-1}\|) \\
&= -\mathcal{P}_{T_{\bar{x}}} A^T A(x_k - x_{k-1}) + \mathcal{P}_{T_{\bar{x}}} A^T(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|) \\
&= -A^T_{T_{\bar{x}}} A_{T_{\bar{x}}}(x_k - x_{k-1}) + A^T_{T_{\bar{x}}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|).
\end{aligned} \tag{B.10}$$

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

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

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

$$A_{T_{\bar{x}}}(x_k - x_{k-1}) = M_{\bar{R}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|), \tag{B.11}$$

which concludes the proof. □

C Trajectory of ADMM

We first provide the fixed-point characterization of ADMM based on the equivalence between ADMM and Douglas–Rachford [19] and Peaceman–Rachford splitting [42] methods, and then present the proofs for the trajectory of ADMM.

C.1 Fixed-point characterization and convergence of ADMM

The dual problem of ($\mathscr{P}_{\text{ADMM}}$) reads

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

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.

C.1.1 Relaxed ADMM and Douglas–Rachford splitting

It is well-known that ADMM is equivalent to applying Douglas–Rachford splitting [19] to the dual problem ($\mathscr{D}_{\text{ADMM}}$). Below we first recall the equivalence between ADMM and Douglas–Rachford which was first established in [25], and then use the convergence of Douglas–Rachford splitting which is well established in the literature [6] to conclude the convergence of ADMM.

For the sake of generality, we consider the following so called *relaxed ADMM*

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

where $\phi \in [0,2]$ is the relaxation parameter. When $\phi = 1$, the relaxed ADMM recovers the standard ADMM (1.2). Below show demonstrate that the relaxed ADMM is equivalent to the relaxed Douglas–Rachford applying to solve ($\mathscr{D}_{\text{ADMM}}$).

- Define $z_k = \psi_k - \gamma(By_k - b)$, we have

$$\begin{aligned}
z_k = \psi_k - \gamma By_k + \gamma b = \psi_{k-1} + \gamma\bar{x}_k &= \phi\psi_{k-1} + \phi\gamma Ax_k + (1-\phi)\psi_{k-1} - (1-\phi)\gamma(By_{k-1} - b)\\
&= (1-\phi)z_{k-1} + \phi(\psi_{k-1} + \gamma Ax_k)\\
&= (1-\phi)z_{k-1} + \phi(z_{k-1} + u_k - \psi_{k-1}).
\end{aligned}$$

  When $\phi = 1$, we have $z_k = \psi_{k-1} + \gamma Ax_k$.
- For the update of x_k, denote $u_k = \psi_{k-1} + \gamma(Ax_k + By_{k-1} - b)$. 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

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

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

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

- Combining all the relations we get

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

  which is exactly the iteration of Douglas–Rachford splitting applied to solve the dual problem ($\mathscr{D}_{\text{ADMM}}$) with .

Define the following operator

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

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

$$z_k = \mathscr{F}_{\text{DR}}(z_{k-1}).$$

It should be noted that for z_k we have $z_k = \psi_k - \gamma B y_k + \gamma b = \psi_{k-1} + \gamma A x_k$ which is the same as in (1.2). Owing to [6], we have that $\mathscr{F}_{\text{DR}}$ is firmly non-expansive with the set of fixed-points $\operatorname{fix}(\mathscr{F}_{\text{DR}})$ being non-empty, and there exists a fixed-point $z^\star \in \operatorname{fix}(\mathscr{F}_{\text{DR}})$ 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 ($\mathscr{D}_{\text{ADMM}}$). 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

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

which are needed in the proofs below.

C.1.2 Symmetric ADMM and Peaceman–Rachford splitting

Below we present a short discussion on the relation between the symmetric ADMM and Peaceman–Rachford splitting method, which was first established in [25].

- For the update of x_k, let $u_k = \psi_{k-\frac{1}{2}} = \psi_{k-1} + \gamma(Ax_k + By_{k-1} - b)$ and $z_k = \psi_k - \gamma B y_k + \gamma b$. As A has full column rank, x_k is the unique minimiser of $R(x) + \frac{\gamma}{2}\|Ax + By_{k-1} - b + \frac{1}{\gamma}\psi_{k-1}\|^2$. Then owing to duality,

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

- For y_k, the full column rank of B ensures the uniqueness of y_k. Since $\psi_k = \psi_{k-\frac{1}{2}} + \gamma(Ax_k + By_k - b)$, then

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

- For z_k, since $u_k = \psi_{k-\frac{1}{2}}$,

$$z_k = \psi_k - \gamma B y_k + \gamma b = u_k + \gamma A x_k = 2u_k - \psi_{k-1} - \gamma(By_{k-1} - b) = z_{k-1} + 2(u_k - \psi_{k-1}).$$

  Combining the above relations we get

$$\begin{aligned}
u_k &= \big(\mathrm{Id} + \gamma\partial(R^* \circ -A^T)\big)^{-1}(2\psi_{k-1} - z_{k-1}),\\
z_k &= z_{k-1} + 2(u_k - \psi_{k-1}),\\
\psi_k &= \big(\mathrm{Id} + \gamma\partial(J^* \circ -B^T)\big)^{-1}(z_k - \gamma b),
\end{aligned} \tag{C.3}$$

  which is the iteration of Peaceman–Rachford splitting when applied to solve ($\mathscr{D}_{\text{ADMM}}$).

Define the following operator

$$\mathscr{F}_{\text{PR}} = \big(2\big(\text{Id}+\gamma\partial(R^*\circ -A^T)\big)^{-1}-\text{Id}\big)\big(2\big(\text{Id}+\gamma\partial(J^*\circ -B^T)\big)^{-1}-\text{Id}\big),$$

then (C.3) can be written as the fixed-point iteration in terms of z_k, that is

$$z_k = \mathscr{F}_{\text{PR}}(z_{k-1}).$$

It should be noted that for z_k we have $z_k = \psi_k - \gamma B y_k + \gamma b = \psi_{k-1} + \gamma A x_k$ which is the same as in (5.4). Different to the case of Douglas–Rachford, the operator $\mathscr{F}_{\text{PR}}$ is only non-expansive [6], hence the conditions for z_k to be convergent is stronger than that of $\mathscr{F}_{\text{DR}}$. However, when it converges, it tends to be faster than Douglas–Rachford splitting [25].

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

Given a saddle point $(x^\star, y^\star, \psi^\star)$ of $\mathscr{L}(x,y;\psi)$, 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

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

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

$$\bar{R}(x) \overset{\text{def}}{=} \frac{1}{\gamma}\big(R(x) - \langle x, -A^T\psi^\star\rangle\big),\ \bar{J}(y) \overset{\text{def}}{=} \frac{1}{\gamma}\big(J(y) - \langle w, -B^T\psi^\star\rangle\big), \tag{C.4}$$

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

$$\begin{aligned} M_{\bar{R}} &\overset{\text{def}}{=} A_R\big(\text{Id} + (A_R^T A_R)^{-1} H_{\bar{R}}\big)^{-1}(A_R^T A_R)^{-1} A_R^T, \\ M_{\bar{J}} &\overset{\text{def}}{=} B_J\big(\text{Id} + (B_J^T B_J)^{-1} H_{\bar{J}}\big)^{-1}(B_J^T B_J)^{-1} B_J^T, \end{aligned} \tag{C.5}$$

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

$$M_{\text{ADMM}} \overset{\text{def}}{=} \frac{1}{2}\text{Id} + \frac{1}{2}(2M_{\bar{R}} - \text{Id})(2M_{\bar{J}} - \text{Id}). \tag{C.6}$$

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

1. Finite manifold identification of ADMM The finite manifold identification of ADMM is already discussed in [38], below we present a short discussion for the sake of self-consistency. At convergence of ADMM, owing to (1.2) we have

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

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

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

and similarly

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

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 [28, Theorem 5.3] are fulfilled for R and J, and there exists K large enough such that for all $k \ge K$, there holds

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

which is the finite manifold identification.

2. linearization 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 $\bar{J}(y) \overset{\text{def}}{=} \beta J(y) - \langle y, -\beta B^T \psi^\star\rangle$ and its Riemannian Hessian $H_{\bar{J}} = \mathcal{P}_{T^J_{y^\star}} \nabla^2_{\mathscr{M}^J_{y^\star}} \bar{J}(y^\star) \mathcal{P}_{T^J_{y^\star}}$. For B, define $B_J = B \circ \mathcal{P}_{T^J_{y^\star}}$, and $M_{\bar{J}} = B_J(\mathrm{Id} + (B_J^T B_J)^{-1} H_{\bar{J}})^{-1}(B_J^T B_J)^{-1} B_J^T$. Then owing to Lemma B.9, we get

$$\begin{aligned} B_J(y_k - y_{k-1}) &= M_{\bar{J}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|) \\ &= -\beta M_{\bar{J}}(z_k - z_{k-1}) + o(\|z_k - z_{k-1}\|). \end{aligned} \tag{C.7}$$

Now consider x_k and let $w_k = \beta(z_{k-1} - 2\psi_{k-1})$, 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 $\bar{R}(x) \overset{\text{def}}{=} \beta R(x) - \langle x, -\beta A^T \psi^\star\rangle$ and its Riemannian Hessian $H_{\bar{R}} = \mathcal{P}_{T^R_{x^\star}} \nabla^2_{\mathscr{M}^R_{x^\star}} \bar{R}(x^\star) \mathcal{P}_{T^R_{x^\star}}$. Denote $A_R = A \circ \mathcal{P}_{T^R_{x^\star}}$, and $M_{\bar{R}} = A_R(\mathrm{Id} + (A_R^T A_R)^{-1} H_{\bar{R}})^{-1}(A_R^T A_R)^{-1} A_R^T$. 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 B.4 is applied. From (C.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 B.9 yields,

$$\begin{aligned} A_R(x_k - x_{k-1}) &= M_{\bar{R}}(w_k - w_{k-1}) + o(\|w_k - w_{k-1}\|) \\ &= -\beta M_{\bar{R}}(z_{k-1} - z_{k-2}) + 2M_{\bar{R}} B_J(y_{k-1} - y_{k-2}) + o(\|z_{k-1} - z_{k-2}\|) \\ &= -\beta M_{\bar{R}}(z_{k-1} - z_{k-2}) + 2\beta M_{\bar{R}} M_{\bar{J}}(z_{k-1} - z_{k-2}) + o(\|z_{k-1} - z_{k-2}\|). \end{aligned} \tag{C.8}$$

Finally, from (1.2), (C.7) and (C.8), we have that

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

which is the desired linearization 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 $M_{\bar{R}}$, 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 $M_{\bar{R}}$,

$$\begin{aligned} M_{\bar{R}} &= A_R\Big((A_R^T A_R)^{-\frac{1}{2}}\big(\mathrm{Id} + (A_R^T A_R)^{-\frac{1}{2}} H_{\bar{R}} (A_R^T A_R)^{-\frac{1}{2}}\big)(A_R^T A_R)^{\frac{1}{2}}\Big)^{-1}(A_R^T A_R)^{-1} A_R^T \\ &= A_R(A_R^T A_R)^{-\frac{1}{2}}\big(\mathrm{Id} + (A_R^T A_R)^{-\frac{1}{2}} H_{\bar{R}} (A_R^T A_R)^{-\frac{1}{2}}\big)^{-1}(A_R^T A_R)^{\frac{1}{2}}(A_R^T A_R)^{-1} A_R^T \\ &= A_R(A_R^T A_R)^{-\frac{1}{2}}\big(\mathrm{Id} + (A_R^T A_R)^{-\frac{1}{2}} H_{\bar{R}} (A_R^T A_R)^{-\frac{1}{2}}\big)^{-1}(A_R^T A_R)^{-\frac{1}{2}} A_R^T. \end{aligned} \tag{C.9}$$

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

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

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

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

Then owing to [6, Example 4.14], $M_{\bar{R}}$ is firmly non-expansive. Similarly, $M_{\bar{J}}$ is firmly non-expansive, and so is M_{ADMM} [6, 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 $M_{\bar{R}}$ and $M_{\bar{J}}$ become

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

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

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

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 $1 \le p \le q \le n$. Consequently, there are $r = p - d$ principal angles $(\zeta_i)_{i=1,\dots,r}$ between T_{A_R} and T_{B_J} that are strictly greater than 0 and smaller than $\pi/2$. Suppose that $\zeta_1 \le \cdots \le \zeta_r$. Define the following two diagonal matrices

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

Owing to [8, 17], there exists a real orthogonal matrix U such that

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

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

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

$$\widetilde{M}_{\text{ADMM}} = U \left[\begin{array}{cc|c} C^2 & CS & 0 \\ -CS & C^2 & 0 \\ \hline 0 & 0 & 0_{n-2r} \end{array}\right] U^T. \tag{C.10}$$

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

$$z_k - z_{k-1} = M_{\text{ADMM}}(z_{k-1} - z_{k-2}) = M_{\text{ADMM}}^k(z_0 - z_{-1}). \tag{C.11}$$

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

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

which means $v_0 \in \ker(M_{\text{ADMM}})$ where $\ker(M_{\text{ADMM}})$ denotes the kernel of M_{ADMM}. Since $M_{\text{ADMM}}^\infty M_{\text{ADMM}}^k = M_{\text{ADMM}}^\infty$, we have $v_k \in \ker(M_{\text{ADMM}})$ holds for any $k \in \mathbb{N}$. Then from (C.11) we have

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

The block diagonal property of (C.10) indicates that there exists an elementary transformation matrix E such that

$$\widetilde{M}_{\text{ADMM}} = UE \begin{bmatrix} B_1 & & & \\ & \ddots & & \\ & & B_r & \\ & & & 0_{n-2r} \end{bmatrix} EU^T,$$

where for each $i = 1, \dots, r$, 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 $i = 1, \dots, d$, 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 $1 \le e < r$, we have

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

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

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

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

$$\begin{aligned} \cos(\theta_k) = \frac{\langle v_k, v_{k-1}\rangle}{\|v_k\|\|v_{k-1}\|} = \frac{\langle u_k, u_{k-1}\rangle}{\|u_k\|\|u_{k-1}\|} &= \frac{\langle s_k + t_k, s_{k-1} + t_{k-1}\rangle}{\|s_k + t_k\|\|s_{k-1} + t_{k-1}\|} \\ &= \frac{\langle s_k, s_{k-1}\rangle}{\|s_k + t_k\|\|s_{k-1} + t_{k-1}\|} + \frac{\langle t_k, t_{k-1}\rangle}{\|s_k + t_k\|\|s_{k-1} + t_{k-1}\|} \\ &= \frac{\|s_k\|^2 \cos(\zeta)}{\|s_k\|^2 + \|t_k\|^2} \cdot \frac{\|s_k + t_k\|}{\|s_{k-1} + t_{k-1}\|} + O(\eta^{2k-1}). \end{aligned} \tag{C.12}$$

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 $\cos(\zeta)$. □

C.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 $M_R \overset{\text{def}}{=} A\big(\mathrm{Id} + \frac{1}{\gamma}(A^TA)^{-1}\nabla^2 R(x^\star)\big)^{-1}(A^TA)^{-1}A^T$, and redefine

$$M_{\text{ADMM}} \overset{\text{def}}{=} \frac{1}{2}\mathrm{Id} + \frac{1}{2}(2M_R - \mathrm{Id})(2M_J - \mathrm{Id}). \tag{C.13}$$

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

1. Linearization 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

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

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 (C.9) 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 $[0, 1]$. □

D Adaptive acceleration for ADMM

D.1 Convergence of A^{3}DMM

Proof of Proposition 4.2. From (4.4), we have that

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

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

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

which means that $\{z_k\}_{k\in\mathbb{N}}$ is quasi-Fejér monotone with respect to $\mathrm{fix}(\mathscr{F})$. Then invoke [6, Proposition 5.34] we obtain the convergence of the sequence $\{z_k\}_{k\in\mathbb{N}}$. □

D.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,

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

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

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

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

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

We prove this by induction. Note that

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

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

$$\begin{aligned} V_k C^{m+1} &\overset{\text{(D.1)}}{=} (MV_{k-1})C^{m+1} \overset{\text{(D.2)}}{=} MV_k C^m + ME_0 C^m = M(V_{m+k} + E_m) + ME_0 C^m \\ &\overset{\text{(D.1)}}{=} V_{m+2} + ME_m + ME_0 C^m \end{aligned}$$

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

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

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

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

We prove this by induction. Note that

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

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

$$\begin{aligned} V_k C^{m+1} &\overset{(D.1)}{=} (MV_{k-1})C^{m+1} \\ &\overset{(D.2)}{=} MV_k C^m + ME_0 C^m = M(V_{m+k} + E_m) + ME_0 C^m \\ &\overset{(D.1)}{=} V_{m+2} + ME_m + ME_0 C^m. \end{aligned}$$

So, plugging in our assumption on E_m, we have

$$E_{m+1} = ME_m + ME_0 C^m = ME_0 C^m + M\left(\sum_{j=1}^{m} M^j E_0 C^{m-j}\right) = \sum_{j=1}^{m+1} M^j E_0 C^{m+1-j}.$$

To bound the extrapolation error,

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

Therefore,

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

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 [49, Section 6.7], 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 $u_k \overset{\text{def}}{=} U v_k$,

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

where $\mathscr{P}_q$ 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

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

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 [49], if $|\lambda_q| > |\lambda_{q+1}|$, then Q has precisely q roots $r_1, \ldots, r_q$ satisfying $r_j = \lambda_j + O(|\lambda_{q+1}/\lambda_j|^k)$. 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

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

Now, $y \overset{\text{def}}{=} \sum_{i=0}^q \gamma_i z_{k-i-1}$ is precisely the MPE update and norm bounds on this are presented in [49]. For completeness, we reproduce their arguments here: Let $A \overset{\text{def}}{=} \mathrm{Id} - M$, 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 $\tilde{\mathscr{P}}_q$, the set of all polynomials of degree q such that $h(1)=1$. Therefore

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

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

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

Finally, in our case where $\lambda(M) = [\alpha,\beta]$ with $1>\beta>\alpha>-1$, it is well known that $\min_{h\in\tilde{\mathscr{P}}_q} \max_{t\in\lambda(M)} |h(t)|$ has an explicit expression (see, for example, [10] or [49, Section 7.3.1]):

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

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

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

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