The Complexity of Primal-Dual Fixed Point Methods for Ridge Regression *

Ademir Alves Ribeiro[§] Peter Richtárik[¶]

August 28, 2017

Abstract

We study the ridge regression (L_2 regularized least squares) problem and its dual, which is also a ridge regression problem. We observe that the optimality conditions describing the primal and dual optimal solutions can be formulated in several different but equivalent ways. The optimality conditions we identify form a linear system involving a structured matrix depending on a single relaxation parameter which we introduce for regularization purposes. This leads to the idea of studying and comparing, in theory and practice, the performance of the fixed point method applied to these reformulations. We compute the optimal relaxation parameters and uncover interesting connections between the complexity bounds of the variants of the fixed point scheme we consider. These connections follow from a close link between the spectral properties of the associated matrices. For instance, some reformulations involve purely imaginary eigenvalues; some involve real eigenvalues and others have all eigenvalues on the complex circle.

We show that the deterministic Quartz method–which is a special case of the randomized dual coordinate ascent method with arbitrary sampling recently developed by Qu, Richtarik and Zhang–can be cast in our framework, and achieves the best rate in theory and in numerical experiments among the fixed point methods we study.

Keywords. Unconstrained minimization, primal-dual methods, ridge regression, gradient descent.

1 Introduction

Given matrices $A_1, \ldots, A_n \in \mathbb{R}^{d\times m}$ encoding n observations (examples), and vectors $y_1, \ldots, y_n \in \mathbb{R}^m$ encoding associated responses (labels), one is often interested in finding a vector $w \in \mathbb{R}^d$ such that, in some precise sense, the product $A_i^T w$ is a good approximation of y_i for all i. A fundamental approach to this problem, used in all areas of computational practice, is to formulate the problem as an L_2-regularized least-squares problem, also known as *ridge regression*. In particular, we consider

*The results of this paper were obtained between October 2014 and March 2015, during AR's affiliation with the University of Edinburgh.

¶School of Mathematics, University of Edinburgh, United Kingdom (email: peter.richtarik@ed.ac.uk), supported by the EPSRC Grant EP/K02325X/1, "Accelerated Coordinate Descent Methods for Big Data Optimization".

§Department of Mathematics, Federal University of Paraná, Brazil (email: ademir.ribeiro@ufpr.br), supported by CNPq, Brazil, Grants 201085/2014-3 and 309437/2016-4.

the *primal* ridge regression problem

$$\min_{w\in\mathbb{R}^d} P(w) \stackrel{\text{def}}{=} \frac{1}{2n}\sum_{i=1}^{n}\|A_i^T w - y_i\|^2 + \frac{\lambda}{2}\|w\|^2 = \frac{1}{2n}\|A^T w - y\|^2 + \frac{\lambda}{2}\|w\|^2, \tag{1}$$

where $\lambda > 0$ is a regularization parameter, $\|\cdot\|$ denotes the standard Euclidean norm. In the second and more concise expression we have concatenated the observation matrices and response vectors to form a single observation matrix $A = [A_1, A_2, \cdots, A_n] \in \mathbb{R}^{d\times N}$ and a single response vector $y = (y_1, y_2, \cdots, y_n) \in \mathbb{R}^N$, where $N = nm$.

With each observation (A_i, y_i) we now associate a dual variable, $\alpha_i \in \mathbb{R}^m$. The Fenchel *dual* of (1) is also a ridge regression problem:

$$\max_{\alpha\in\mathbb{R}^N} D(\alpha) \stackrel{\text{def}}{=} -\frac{1}{2\lambda n^2}\|A\alpha\|^2 + \frac{1}{n}\alpha^T y - \frac{1}{2n}\|\alpha\|^2, \tag{2}$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{R}^N$.

Optimality conditions. The starting point of this work is the observation that the *optimality conditions* for the primal and dual ridge regression problems can be written in several *different* ways, in the form of a linear system involving the primal and dual variables. In particular, we find several different matrix-vector pairs (M, b), where $M \in \mathbb{R}^{(d+N)\times(d+N)}$ and $b \in \mathbb{R}^{d+N}$, such that the optimality conditions can be expressed in the form of a linear system as

$$x = Mx + b, \tag{3}$$

where $x = (w, \alpha) \in \mathbb{R}^{d+N}$.

Fixed point methods. With each system (3) one can naturally associate a fixed point method performing the iteration $x^{k+1} = Mx^k + b$. However, unless the spectrum of M is contained in the unit circle, such a method will not converge [**?**]. To overcome this drawback, we utilize the idea of *relaxation*. In particular, we pick a relaxation parameter $\theta \neq 0$ and replace (3) with the equivalent system

$$x = G_\theta x + b_\theta,$$

where $G_\theta = (1-\theta)I + \theta M$ and $b_\theta = \theta b$. The choice $\theta = 1$ recovers (3). We then study the convergence of the *primal-dual fixed point methods*

$$x^{k+1} = G_\theta x^k + b_\theta$$

through a careful study of the spectra of the iteration matrices G_θ.

The central question of this work is:

While all these formulations are necessarily algebraically equivalent, they give rise to different fixed-point algorithms.

1.1 Contributions and literature review

It is well known that the role of duality in optimization and machine learning is very important, not only from the theoretical point of view but also computationally [10, 11, 15].

However, a more recent idea that has generated many contributions is the usage of the primal and dual problems together. Primal-dual methods have been employed in convex optimization problems where strong duality holds, obtaining success when applied to several types of nonlinear and nonsmooth functions that arise in various application fields, such as image processing, machine learning, inverse problems, among others [2, 7, 8].

On the other hand, fixed-point-type algorithms are classical tools for solving some structured linear systems. In particular, we have the iterative schemes developed by the mathematical economists Arrow, Hurwicz and Uzawa for solving saddle point problems [1, 13].

In this paper we develop several primal-dual fixed point methods for the Ridge Regression problem. Ridge regression was introduced by Hoerl and Kennard [5, 6] as a regularization method for solving least squares problems with highly correlated predictors. The goal is to reduce the standard errors of regression coefficients by imposing a penalty, in the L_2 norm, on their size.

Since then, numerous papers were devoted to the study of ridge regression or even for solving problems with a general formulation in which ridge regression is a particular case. Some of these works have considered its dual formulation, proposing deterministic and stochastic algorithms that can be applied to the dual problem [3, 4, 9, 10, 14, 15].

To the best of our knowledge, the only work that considers a primal-dual fixed point approach to deal with ridge regression is [12], where the authors deal with ill-conditioned problems. They present an algorithm based on the gradient method and an accelerated version of this algorithm.

Here we propose methods based on the optimality conditions for the problem of minimizing the duality gap between the ridge regression problems (1) and (2) in different and equivalent ways by means of linear systems involving structured matrices. We also study the complexity of the proposed methods and prove that our main method achieves the optimal accelerated Nesterov rate.

1.2 Outline

In Section 2 we formulate the optimality conditions for the problem of minimizing the duality gap between (1) and (2) in two different, but equivalent, ways by means of linear systems involving structured matrices. We also establish the duality relationship between the problems (1) and (2). In Section 3 we describe a family of (parameterized) fixed point methods applied to the reformulations for the optimality conditions. We present the convergence analysis and complexity results for these methods. Section 4 brings the main contribution of this work, with an accelerated version of the methods described in Section 3. In Section 5 we discuss some variants of our accelerated algorithm. Finally, in Section 6 we perform some numerical experiments and present concluding remarks. We left the proofs of the results of this paper to the appendix.

2 Separable and Coupled Optimality Conditions

Defining $x = (w, \alpha) \in \mathbb{R}^{d+N}$, our primal-dual problem consists of minimizing the duality gap between the problems (1) and (2), that is

$$\min_{x \in \mathbb{R}^{d+N}} f(x) \stackrel{\text{def}}{=} P(w) - D(\alpha). \tag{4}$$

This is a quadratic strongly convex problem and therefore admits a unique global solution $x^* \in \mathbb{R}^{d+N}$.

2.1 A separable system

Note that $\nabla f(x) = \begin{pmatrix} \nabla P(w) \\ -\nabla D(\alpha) \end{pmatrix}$, where

$$\nabla P(w) = \frac{1}{n} A(A^T w - y) + \lambda w \quad \text{and} \quad \nabla D(\alpha) = -\frac{1}{\lambda n^2} A^T A \alpha - \frac{1}{n}\alpha + \frac{1}{n} y. \tag{5}$$

So, the first and natural way of writing the optimality conditions for problem (4) is just to set the expressions given in (5) equal to zero, which can be written as

$$\boxed{\begin{pmatrix} w \\ \alpha \end{pmatrix} = -\frac{1}{\lambda n} \begin{pmatrix} AA^T & 0 \\ 0 & A^T A \end{pmatrix} \begin{pmatrix} w \\ \alpha \end{pmatrix} + \frac{1}{\lambda n} \begin{pmatrix} Ay \\ \lambda n y \end{pmatrix}.} \tag{6}$$

2.2 A coupled system

In order to derive the duality between (1) and (2), as well as to reformulate the optimality conditions for problem (4), note that

$$P(w) = \frac{1}{n} \sum_{i=1}^{n} \phi_i(A_i^T w) + \lambda g(w), \tag{7}$$

where $\phi_i(z) = \frac{1}{2}\|z - y_i\|^2$ and $g(w) = \frac{1}{2}\|w\|^2$.

Now, recall that the Fenchel conjugate of a convex function $\xi : \mathbb{R}^l \to \mathbb{R}$ is $\xi^* : \mathbb{R}^l \to \mathbb{R} \cup \{\infty\}$ defined by

$$\xi^*(u) \stackrel{\text{def}}{=} \sup_{s \in \mathbb{R}^l} \{s^T u - \xi(s)\}.$$

Note that if ξ is strongly convex, then $\xi^*(u) < \infty$ for all $u \in \mathbb{R}^l$. Indeed, in this case ξ is bounded below by a strongly convex quadratic function, implying that the "sup" above is in fact a "max".

It is easily seen that $\phi_i^*(s) = \frac{1}{2}\|s\|^2 + s^T y_i$ and $g^*(u) = \frac{1}{2}\|u\|^2$. Furthermore, we have

$$D(\alpha) = -\lambda g^* \left(\frac{1}{\lambda n} \sum_{i=1}^{n} A_i \alpha_i \right) - \frac{1}{n} \sum_{i=1}^{n} \phi_i^*(-\alpha_i). \tag{8}$$

If we write

$$\bar{\alpha} \stackrel{\text{def}}{=} \frac{1}{\lambda n} A\alpha = \frac{1}{\lambda n} \sum_{i=1}^{n} A_i \alpha_i, \tag{9}$$

the duality gap can be written as

$$P(w) - D(\alpha) = \lambda\big(g(w) + g^*(\bar{\alpha}) - w^T \bar{\alpha}\big) + \frac{1}{n} \sum_{i=1}^{n} \Big(\phi_i(A_i^T w) + \phi_i^*(-\alpha_i) + \alpha_i^T A_i^T w\Big)$$

and the weak duality follows immediately from the fact that

$$g(w) + g^*(\bar{\alpha}) - w^T \bar{\alpha} \geq 0 \quad \text{and} \quad \phi_i(A_i^T w) + \phi_i^*(-\alpha_i) + \alpha_i^T A_i^T w \geq 0.$$

Strong duality occurs when these quantities vanish, which is precisely the same as

$$w = \nabla g^*(\bar{\alpha}) \quad \text{and} \quad \alpha_i = -\nabla \phi_i(A_i^T w)$$

or, equivalently,

$$\bar{\alpha} = \nabla g(w) \quad \text{and} \quad A_i^T w = \nabla \phi_i^*(-\alpha_i).$$

Therefore, another way to see the optimality conditions for problem (4) is by the relations

$$w = \bar{\alpha} = \frac{1}{\lambda n} A\alpha \quad \text{and} \quad \alpha = y - A^T w. \tag{10}$$

This is equivalent to

$$\boxed{\begin{pmatrix} w \\ \alpha \end{pmatrix} = -\frac{1}{\lambda n} \begin{pmatrix} 0 & -A \\ \lambda n A^T & 0 \end{pmatrix} \begin{pmatrix} w \\ \alpha \end{pmatrix} + \begin{pmatrix} 0 \\ y \end{pmatrix}.} \tag{11}$$

2.3 Compact form

Both reformulations of the optimality conditions, (6) and (11), can be viewed in the compact form

$$x = Mx + b, \tag{12}$$

for some $M \in \mathbb{R}^{(d+N)\times(d+N)}$ and $b \in \mathbb{R}^{d+N}$. Let us denote

$$M_1 = -\frac{1}{\lambda n} \begin{pmatrix} AA^T & 0 \\ 0 & A^T A \end{pmatrix} \quad \text{and} \quad M_2 = -\frac{1}{\lambda n} \begin{pmatrix} 0 & -A \\ \lambda n A^T & 0 \end{pmatrix} \tag{13}$$

the matrices associated with the optimality conditions formulated as (6) and (11), respectively. Also, let

$$b_1 = \frac{1}{\lambda n} \begin{pmatrix} Ay \\ \lambda n y \end{pmatrix} \quad \text{and} \quad b_2 = \begin{pmatrix} 0 \\ y \end{pmatrix}. \tag{14}$$

Thus, we can rewritte (6) and (11) as

$$x = M_1 x + b_1 \quad \text{and} \quad x = M_2 x + b_2, \tag{15}$$

respectively.

3 Primal-Dual Fixed Point Methods

A method that arises immediately from the relation (12) is given by the scheme

$$x^{k+1} = Mx^k + b.$$

However, unless the spectrum of M is contained in the unit circle, this scheme will not converge. To overcome this drawback, we utilize the idea of *relaxation*. More precisely, we consider a relaxation parameter $\theta \neq 0$ and replace (12) with the equivalent system

$$x = (1-\theta)x + \theta(Mx + b).$$

Note that the choice $\theta = 1$ recovers (12).

The proposed algorithm is then given by the following framework.

Algorithm 3.1 (Primal-Dual Fixed Point Method)

INPUT: matrix $M \in \mathbb{R}^{(d+N)\times(d+N)}$, vector $b \in \mathbb{R}^{d+N}$, parameter $\theta > 0$
STARTING POINT: $x^0 \in \mathbb{R}^{d+N}$
REPEAT FOR $k = 0, 1, 2, \ldots$
    Set $x^{k+1} = (1-\theta)x^k + \theta(Mx^k + b)$

As we shall see later, the use of the relaxation parameter θ enables us to prove convergence of Algorithm 3.1 with $M = M_1$ and $b = b_1$ or $M = M_2$ and $b = b_2$, chosen according to (13) and (14), independent of the spectral radius of these matrices.

Let us denote

$$G(\theta) = (1-\theta)I + \theta M \tag{16}$$

and let x^* be the solution of the problem (4). Then $x^* = Mx^* + b$ with $M = M_1$ and $b = b_1$ or $M = M_2$ and $b = b_2$. Therefore,

$$x^* = G(\theta)x^* + \theta b.$$

Further, the iteration of Algorithm 3.1 can be written as $x^{k+1} = G(\theta)x^k + \theta b$. Thus,

$$\|x^k - x^*\| \leq \|G(\theta)^k\|\|x^0 - x^*\| \tag{17}$$

and consequently the convergence of the algorithm depends on the spectrum of $G(\theta)$. More precisely, it converges if the spectral radius of $G(\theta)$ is less than 1, because in this case we have $G(\theta)^k \to 0$.

In fact, we will address the following questions:

- What is the range for θ so that this scheme converges?
- What is the best choice of θ?
- What is the rate of convergence?
- How the complexity of this algorithm compares with the known ones?

3.1 Convergence analysis

In this section we study the convergence of Algorithm 3.1 and answer the questions raised above. To this end we point out some properties of the iteration matrices and uncover interesting connections between the complexity bounds of the variants of the fixed point scheme we consider. These connections follow from a close link between the spectral properties of the associated matrices.

For this purpose, let

$$A = U\Sigma V^T \tag{18}$$

be the singular value decomposition of A. That is, $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{N\times N}$ are orthogonal matrices and

$$\Sigma = \begin{pmatrix} \widetilde{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \begin{matrix} p \\ d-p \end{matrix} \tag{19}$$
$$\begin{matrix} p & N-p \end{matrix}$$

where $\widetilde{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ brings the (nonzero) singular values $\sigma_1 \geq \cdots \geq \sigma_p > 0$ of A.

The next result is crucial for the convergence analysis and complexity study of Algorithm 3.1.

Lemma 3.1 *The characteristic polynomials of the matrices M_1 and M_2, defined in* (13)*, are*

$$p_1(t) = t^{N+d-2p} \prod_{j=1}^{p} \left(t + \frac{1}{\lambda n} \sigma_j^2 \right)^2 \quad \text{and} \quad p_2(t) = t^{N+d-2p} \prod_{j=1}^{p} \left(t^2 + \frac{1}{\lambda n} \sigma_j^2 \right),$$

respectively.

The following result follows directly from Lemma 3.1 and the fact that M_1 is symmetric.

Corollary 3.2 *The spectral radii of M_1 and M_2 are, respectively,*

$$\rho_1 = \|M_1\| = \frac{\sigma_1^2}{\lambda n} = \frac{\|A\|^2}{\lambda n} \quad \text{and} \quad \rho_2 = \frac{\sigma_1}{\sqrt{\lambda n}} = \frac{\|A\|}{\sqrt{\lambda n}}.$$

From Corollary 3.2 we conclude that if $\sigma_1 < \sqrt{\lambda n}$, then $\rho_1 \leq \rho_2 < 1$. So, $M_1^k \to 0$ and $M_2^k \to 0$, which in turn implies that the pure fixed point method, that is, Algorithm 3.1 with $\theta = 1$, converges. However, if $\sigma_1 \geq \sqrt{\lambda n}$, we cannot guarantee convergence of the pure method.

Now we shall see that Algorithm 3.1 converges for a broad range of the parameter θ, without any assumption on σ_1, λ or n. We begin with the analysis of the framework that uses M_1 and b_1, defined in (13) and (14).

3.2 Fixed Point Method based on M_1

Algorithm 3.2 (Primal-Dual Fixed Point Method; $M = M_1$)

INPUT: $M = M_1$, $b = b_1$, parameter $\theta > 0$
STARTING POINT: $x^0 \in \mathbb{R}^{d+N}$
REPEAT FOR $k = 0, 1, 2, \ldots$
    Set $x^{k+1} = (1 - \theta)x^k + \theta(Mx^k + b)$

Theorem 3.3 *Let $x^0 \in \mathbb{R}^{d+N}$ be an arbitrary starting point and consider the sequence $(x^k)_{k \in \mathbb{N}}$ generated by Algorithm 3.2 with $\theta \in \left(0, \frac{2\lambda n}{\lambda n + \sigma_1^2}\right)$. Then the sequence (x^k) converges to the (unique) solution of the problem* (4) *at a linear rate of*

$$\rho_1(\theta) \stackrel{\text{def}}{=} \max \left\{ \left| 1 - \theta \left(1 + \frac{\sigma_1^2}{\lambda n} \right) \right|, 1 - \theta \right\}$$

Furthermore, if we choose $\theta_1^* \stackrel{\text{def}}{=} \frac{2\lambda n}{2\lambda n + \sigma_1^2}$, then the (theoretical) convergence rate is optimal and it is equal to

$$\rho_1^* \stackrel{\text{def}}{=} \frac{\sigma_1^2}{2\lambda n + \sigma_1^2} = 1 - \theta_1^*.$$

The top picture of Figure 1 illustrates the eigenvalues of $G_1(\theta) \stackrel{\text{def}}{=} (1-\theta)I + \theta M_1$ (magenta) together with the eigenvalues of M_1 (blue), for a fixed value of the parameter θ. The one farthest from the origin is $1 - \theta - \dfrac{\theta\sigma_1^2}{\lambda n}$ or $1-\theta$. On the bottom we show the two largest (in absolute value) eigenvalues of $G_1(\theta)$ corresponding to the optimal choice of θ.

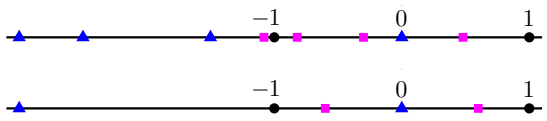

Figure 1: Eigenvalues of $G_1(\theta)$ and M_1.

Now we analyze the fixed point framework that employs M_2 and b_2, defined in (13) and (14).

3.3 Fixed Point Method based on M_2

Algorithm 3.3 (Primal-Dual Fixed Point Method; $M = M_2$)

INPUT: $M = M_2$, $b = b_2$, parameter $\theta > 0$
STARTING POINT: $x^0 \in \mathbb{R}^{d+N}$
REPEAT FOR $k = 0, 1, 2, \ldots$
    Set $x^{k+1} = (1-\theta)x^k + \theta(Mx^k + b)$

Theorem 3.4 *Let $x^0 \in \mathbb{R}^{d+N}$ be an arbitrary starting point and consider the sequence $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm 3.3 with $\theta \in \left(0, \dfrac{2\lambda n}{\lambda n + \sigma_1^2}\right)$. Then the sequence (x^k) converges to the (unique) solution of the problem (4) at an asymptotic convergence rate of*

$$\rho_2(\theta) \stackrel{\text{def}}{=} \sqrt{(1-\theta)^2 + \frac{\theta^2\sigma_1^2}{\lambda n}}.$$

Furthermore, if we choose $\theta_2^* \stackrel{\text{def}}{=} \dfrac{\lambda n}{\lambda n + \sigma_1^2}$, then the (theoretical) convergence rate is optimal and it is equal to

$$\rho_2^* \stackrel{\text{def}}{=} \frac{\sigma_1}{\sqrt{\lambda n + \sigma_1^2}} = \sqrt{1 - \theta_2^*}.$$

The left picture of Figure 2 illustrates, in the complex plane, the eigenvalues of $G_2(\theta) \stackrel{\text{def}}{=} (1-\theta)I + \theta M_2$ (magenta) together with the eigenvalues of M_2 (blue), for a fixed value of the parameter θ. On the right we show, for each $\theta \in (0,1)$, one of the two eigenvalues of $G_2(\theta)$ farthest from the origin. The dashed segment corresponds to the admissible values for θ, that is, the eigenvalues with modulus less than one. The square corresponds to the optimal choice of θ.

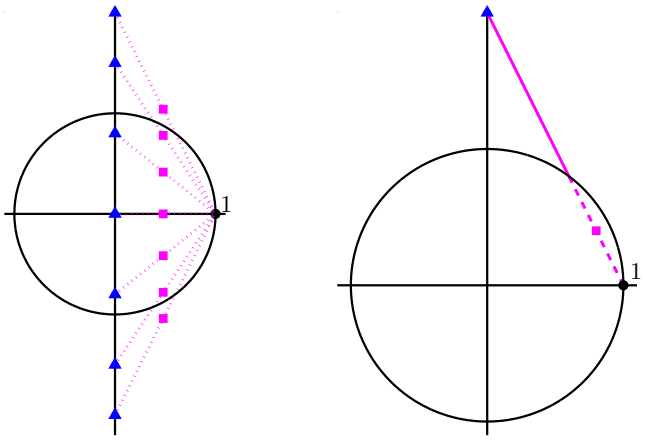

Figure 2: Eigenvalues of $G_2(\theta)$ and M_2.

3.4 Comparison of the rates

We summarize the discussion above in Table 1 which brings the comparison between the pure ($\theta = 1$) and optimal ($\theta = \theta_j^*$, $j = 1, 2$) versions of Algorithm 3.1. We can see that the convergence rate of the optimal version is $\frac{\lambda n}{2\lambda n + \sigma_1^2}$ times the one of the pure version if M_1 is employed and $\sqrt{\frac{\lambda n}{\lambda n + \sigma_1^2}}$ times the pure version when using M_2. Moreover, in any case, employing M_1 provides faster convergence. This can be seen in Figure 3, where Algorithm 3.1 was applied to solve the problem (4). The dimensions considered were $d = 200$, $m = 1$ and $n = 5000$ (so that the total dimension is $d + N = d + nm = 5200$).

We also mention that the pure version does not require the knowledge of σ_1, but it may not converge. On the other hand, the optimal version always converges, but θ depends on σ_1.

	MFP(M_1, θ)	MFP(M_2, θ)
Range for θ	$\left(0, \dfrac{2\lambda n}{\lambda n + \sigma_1^2}\right)$	$\left(0, \dfrac{2\lambda n}{\lambda n + \sigma_1^2}\right)$
Pure ($\theta = 1$)	$\dfrac{\sigma_1^2}{\lambda n}$	$\dfrac{\sigma_1}{\sqrt{\lambda n}}$
Optimal ($\theta = \theta_j^*$)	$\dfrac{\sigma_1^2}{2\lambda n + \sigma_1^2} = 1 - \theta_1^*$	$\sqrt{\dfrac{\sigma_1^2}{\lambda n + \sigma_1^2}} = \sqrt{1 - \theta_2^*}$

Table 1: Comparison between the convergence rates of pure and optimal versions of Algorithm 3.1.

3.5 Direct relationship between the iterates of the two methods

Another relation regarding the employment of M_1 or M_2 in the pure version of Algorithm 3.1, which is also illustrated in Figure 3, is that one step of the method with M_1 corresponds exactly to

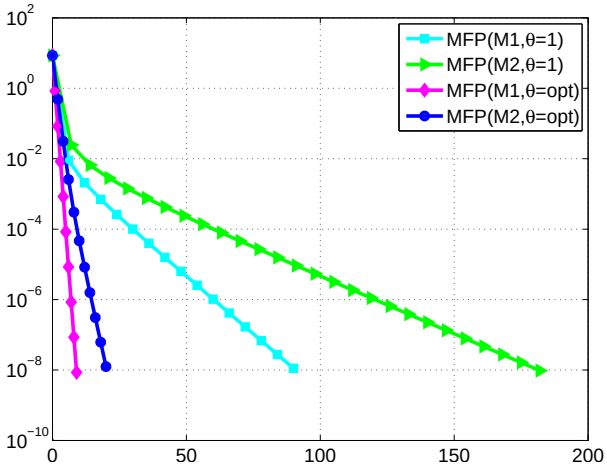

Figure 3: Performance of pure and optimal versions of Algorithm 3.1.

two steps of the one with M_2. Indeed, note first that $M_2^2 = M_1$. Thus, denoting the current point by x and the next iterate by x_M^+, in view of (15) we have

$$\begin{aligned} x_{M_2}^{++} &= M_2 x_{M_2}^+ + b_2 \\ &= M_2(M_2 x + b_2) + b_2 \\ &= M_1 x + M_2 b_2 + b_2 \\ &= M_1 x + b_1 \\ &= x_{M_1}^+. \end{aligned}$$

In Section 4 we shall see how this behavior can invert with a small change in the computation of the dual variable.

3.6 Complexity results

In order to establish the complexity of Algorithm 3.1 we need to calculate the condition number of the objective function, defined in (4). Note that the Hessian of f is given by

$$\nabla^2 f = \frac{1}{n} \begin{pmatrix} AA^T + \lambda n I & 0 \\ 0 & \frac{1}{\lambda n} A^T A + I \end{pmatrix}$$

Let us consider two cases:

- If $\lambda n \geq 1$, then $\sigma_1^2 + \lambda n \geq \sigma_1^2 + 1 \geq \dfrac{\sigma_1^2}{\lambda n} + 1$, which in turn implies that the largest eigenvalue

of $\nabla^2 f$ is $L = \dfrac{\sigma_1^2 + \lambda n}{n}$. The smallest eigenvalue is

$$\begin{cases} \dfrac{1}{n}, \text{ if } d < N \\ \dfrac{\sigma_d^2}{\lambda n^2} + \dfrac{1}{n}, \text{ if } d = N \\ \min\left\{\lambda, \dfrac{\sigma_N^2}{\lambda n^2} + \dfrac{1}{n}\right\}, \text{ if } d > N. \end{cases}$$

Therefore, if $d < N$, the condition number of $\nabla^2 f$ is the condition number of $\nabla^2 f$ is

$$\sigma_1^2 + \lambda n. \tag{20}$$

- If $\lambda n < 1$, then $\sigma_1^2 + \lambda n < \sigma_1^2 + 1 < \dfrac{\sigma_1^2}{\lambda n} + 1$, which in turn implies that the largest eigenvalue of $\nabla^2 f$ is $L = \dfrac{\sigma_1^2 + \lambda n}{\lambda n^2}$. The smallest eigenvalue is

$$\begin{cases} \min\left\{\dfrac{\sigma_d^2}{n} + \lambda, \dfrac{1}{n}\right\}, \text{ if } d < N \\ \dfrac{\sigma_d^2}{n} + \lambda, \text{ if } d = N \\ \lambda, \text{ if } d > N. \end{cases}$$

So, assuming that $d < N$, the condition number is $\dfrac{\sigma_1^2 + \lambda n}{\lambda n^2 \min\left\{\dfrac{\sigma_d^2}{n} + \lambda, \dfrac{1}{n}\right\}}$. If A is rank deficient, then the condition number is

$$\frac{\sigma_1^2 + \lambda n}{(\lambda n)^2}. \tag{21}$$

We stress that despite the analysis was made in terms of the sequence $x^k = (w^k, \alpha^k)$, the linear convergence also applies to objective values. Indeed, since f is L-smooth, we have

$$f(x^k) \leq f(x^*) + \nabla f(x^*)^T (x^k - x^*) + \frac{L}{2}\|x^k - x^*\|^2 = \frac{L}{2}\|x^k - x^*\|^2,$$

where the equality follows from the fact that the optimal objective value is zero. Therefore, if we want to get $f(x^k) - f(x^*) < \varepsilon$ and we have linear convergence rate ρ on the sequence (x^k), then it is enough to enforce

$$\frac{L}{2}\rho^{2k}\|x^0 - x^*\|^2 < \varepsilon,$$

or equivalently,

$$k > \frac{-1}{2\log\rho} \log\left(\frac{\|x^0 - x^*\|^2 L}{2\varepsilon}\right). \tag{22}$$

Using the estimate $\log(1-\theta) \approx -\theta$, we can approximate the second hand side of (22) by

$$\frac{1}{2\theta_1^*} \log\left(\frac{\|x^0 - x^*\|^2 L}{2\varepsilon}\right), \tag{23}$$

in the case M_1 is used and by

$$\frac{1}{\theta_2^*} \log\left(\frac{\|x^0 - x^*\|^2 L}{2\varepsilon}\right), \tag{24}$$

if we use M_2.

In order to estimate the above expressions in terms of the condition number, let us consider the more common case $\lambda n \geq 1$. Then the condition number of $\nabla^2 f$ is given by (20), that is,

$$\kappa \stackrel{\text{def}}{=} \sigma_1^2 + \lambda n. \tag{25}$$

So, if we use M_1, the complexity is proportional to

$$\frac{1}{2\theta_1^*} = \frac{\sigma_1^2 + 2\lambda n}{4\lambda n} = \frac{\kappa + \lambda n}{4\lambda n}. \tag{26}$$

If we use M_2, the complexity is proportional to

$$\frac{1}{\theta_2^*} = \frac{\lambda n + \sigma_1^2}{\lambda n} = \frac{\kappa}{\lambda n}. \tag{27}$$

4 Accelerated Primal-Dual Fixed Point Method

Now we present our main contribution. When we employ Algorithm 3.3, the primal and dual variables are mixed in two equations. More precisely, in view of (9) the iteration in this case can be rewritten as

$$\begin{cases} w^{k+1} = (1-\theta)w^k + \theta\bar{\alpha}^k \\ \alpha^{k+1} = (1-\theta)\alpha^k + \theta(y - A^T w^k). \end{cases}$$

The idea here is to apply block Gauss-Seidel to this system. That is, we use the freshest w to update α. Let us state formally the method by means of the following framework.

Algorithm 4.1 (Accelerated Fixed Point Method)

INPUT: matrix $A \in \mathbb{R}^{d\times N}$, vector $y \in \mathbb{R}^N$, parameter $\theta \in (0,1]$
STARTING POINTS: $w^0 \in \mathbb{R}^d$ and $\alpha^0 \in \mathbb{R}^N$
REPEAT FOR $k = 0, 1, 2, \ldots$
    Set $w^{k+1} = (1-\theta)w^k + \theta\bar{\alpha}^k$
    Set $\alpha^{k+1} = (1-\theta)\alpha^k + \theta(y - A^T w^{k+1})$

Due to this modification, we can achieve faster convergence. This algorithm is a deterministic version of a randomized primal-dual algorithm (Quartz) proposed and analyzed by Qu, Richtárik and Zhang [8].

4.1 Convergence analysis

In this section we study the convergence of Algorithm 4.1. We shall determine all values for the parameter θ for which this algorithm converges as well as the one giving the best convergence rate.

To this end, we start by showing that Algorithm 4.1 can be viewed as a fixed point scheme. Then we determine the "dynamic" spectral properties of the associated matrices, which are parameterized by θ.

First, note that the iteration of our algorithm can be written as

$$\begin{pmatrix} I & 0 \\ \theta A^T & I \end{pmatrix} \begin{pmatrix} w^{k+1} \\ \alpha^{k+1} \end{pmatrix} = \begin{pmatrix} (1-\theta)I & \frac{\theta}{\lambda n} A \\ 0 & (1-\theta)I \end{pmatrix} \begin{pmatrix} w^k \\ \alpha^k \end{pmatrix} + \begin{pmatrix} 0 \\ \theta y \end{pmatrix}$$

or in a compact way as

$$x^{k+1} = G_3(\theta) x^k + f \tag{28}$$

with

$$G_3(\theta) = \begin{pmatrix} I & 0 \\ \theta A^T & I \end{pmatrix}^{-1} \begin{pmatrix} (1-\theta)I & \frac{\theta}{\lambda n} A \\ 0 & (1-\theta)I \end{pmatrix} = (1-\theta)I + \theta \begin{pmatrix} 0 & \frac{1}{\lambda n} A \\ (\theta-1)A^T & -\frac{\theta}{\lambda n} A^T A \end{pmatrix} \tag{29}$$

and

$$f = \begin{pmatrix} I & 0 \\ \theta A^T & I \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \theta y \end{pmatrix} = \begin{pmatrix} 0 \\ \theta y \end{pmatrix}.$$

We know that if the spectral radius of $G_3(\theta)$ is less that 1, then the sequence defined by (28) converges. Indeed, in this case the limit point is just x^*, the solution of the problem (4). This follows from the fact that $x^* = G_3(\theta)x^* + f$.

Next lemma provides the spectrum of $G_3(\theta)$.

Lemma 4.1 *The eigenvalues of the matrix $G_3(\theta)$, defined in (29), are given by*

$$\frac{1}{2\lambda n}\left\{2(1-\theta)\lambda n - \theta^2\sigma_j^2 \pm \theta\sigma_j\sqrt{\theta^2\sigma_j^2 - 4(1-\theta)\lambda n}\,, \quad j = 1, \ldots, p\right\} \cup \{1-\theta\}.$$

Figure 4 illustrates, in the complex plane, the spectrum of the matrix $G_3(\theta)$ for many different values of θ. We used $n = 250$, $d = 13$, $m = 1$ (therefore $N = 250$), $\lambda = 0.3$ and a random matrix $A \in \mathbb{R}^{d\times N}$. The pictures point out the fact that for some range of θ the spectrum is contained in a circle and for other values of θ some of the eigenvalues remain in a circle while others are distributed along the real line, moving monotonically as this parameter changes. These statements will be proved in the sequel.

In what follows, let us consider the functions $\delta_j : [0,1] \to \mathbb{R}$ defined by

$$\delta_j(\theta) = \theta^2\sigma_j^2 - 4(1-\theta)\lambda n. \tag{30}$$

The following result brings some basic properties of them, illustrated in Figure 5.

Lemma 4.2 *Each function δ_j, $j = 1, \ldots, p$, is strictly increasing, from $-4\lambda n$ to σ_j^2 as θ goes from zero to 1. Furthermore, these functions are sorted in decreasing order, $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_p$, and their zeros,*

$$\bar{\theta}_j \stackrel{\text{def}}{=} \frac{-2\lambda n + 2\sqrt{\lambda n(\lambda n + \sigma_j^2)}}{\sigma_j^2}, \tag{31}$$

are sorted in increasing order: $0 < \bar{\theta}_1 \leq \bar{\theta}_2 \leq \cdots \leq \bar{\theta}_p < 1$.

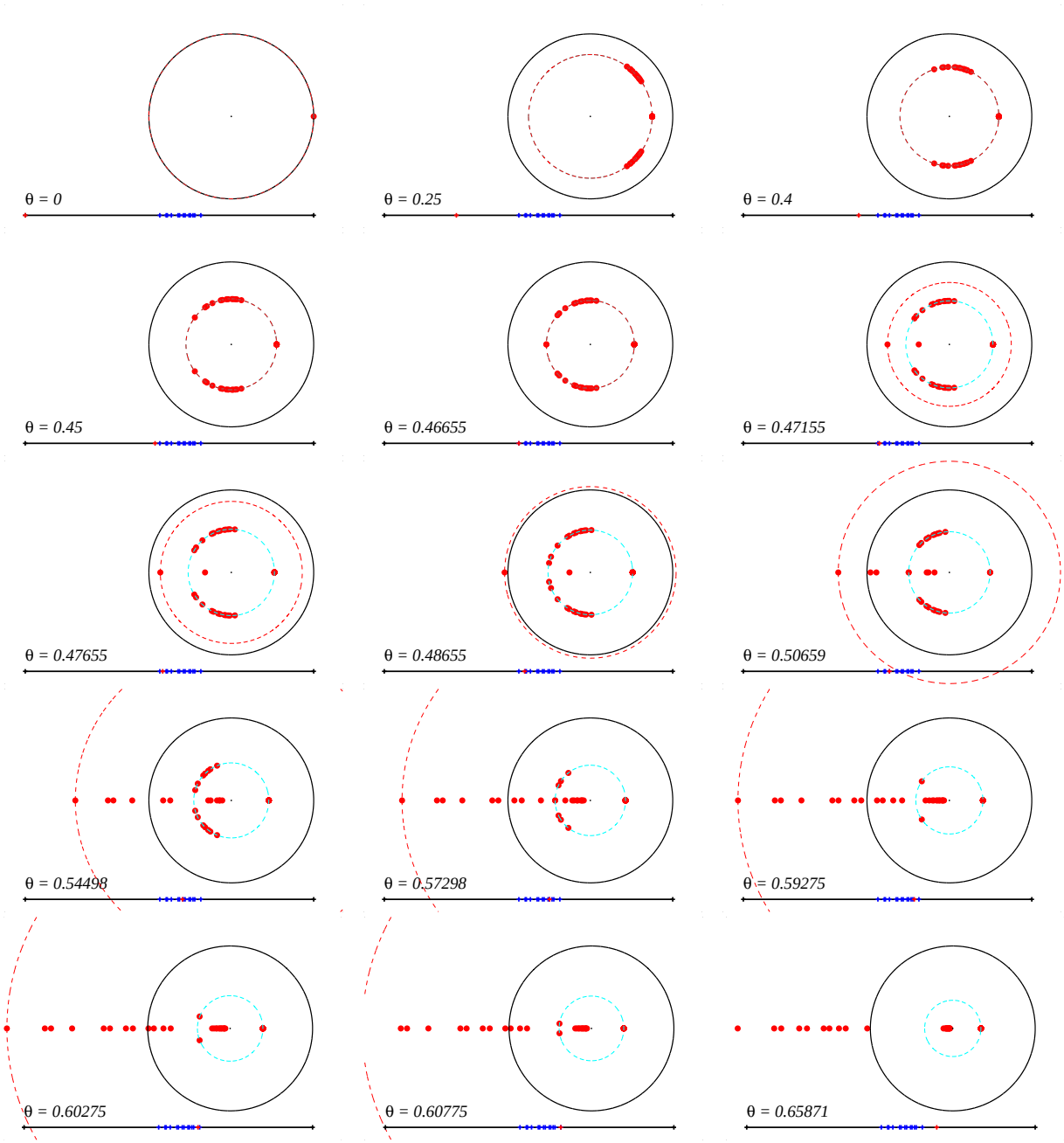

Figure 4: The spectrum of $G_3(\theta)$ for many different values of θ. The first 4 pictures satisfying the condition $\theta^2\sigma_1^2 - 4(1-\theta)\lambda n < 0$; in the fifth picture we have $\theta^2\sigma_1^2 - 4(1-\theta)\lambda n = 0$ and the remaining ones represent the case where $\theta^2\sigma_1^2 - 4(1-\theta)\lambda n > 0$.

Proof. It is straightforward. □

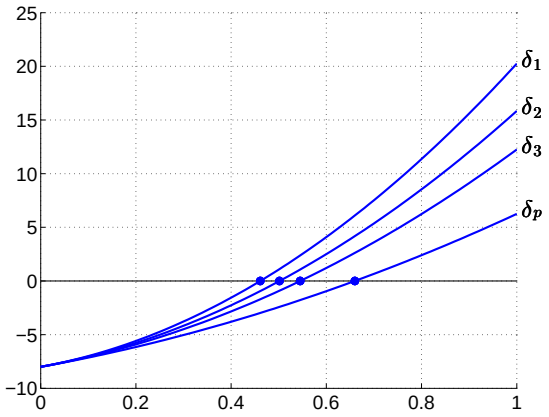

Figure 5: The functions δ_j, $j = 1, \ldots, p$ and the properties stated in Lemma 4.2.

Now we shall study the spectrum of $G_3(\theta)$, given in Lemma 4.1. For this, let us denote

$$\lambda_0(\theta) \stackrel{\text{def}}{=} (1 - \theta) \tag{32}$$

and, for $j = 1, \ldots, p$,

$$\lambda_j^-(\theta) \stackrel{\text{def}}{=} \frac{1}{2\lambda n}\left(2(1-\theta)\lambda n - \theta^2\sigma_j^2 - \theta\sigma_j\sqrt{\delta_j(\theta)}\right), \tag{33}$$

$$\lambda_j^+(\theta) \stackrel{\text{def}}{=} \frac{1}{2\lambda n}\left(2(1-\theta)\lambda n - \theta^2\sigma_j^2 + \theta\sigma_j\sqrt{\delta_j(\theta)}\right) \tag{34}$$

where δ_j is defined in (30).

Lemma 4.3 *Consider $\bar{\theta}_1$ as defined in* (31)*. If $\theta \in [0, \bar{\theta}_1]$, then*

$$|\lambda_j^+(\theta)| = |\lambda_j^-(\theta)| = 1 - \theta$$

for all $j = 1, \ldots, p$, which in turn implies that the spectral radius of $G_3(\theta)$ is $1 - \theta$.

It can be shown that the parameter $\theta = \theta_1^*$, defined in Theorem 3.3, satisfies the conditions of Lemma 4.3. So, the spectral radius of $G_3(\theta_1^*)$ is $1 - \theta_1^*$, exactly the same spectral radius of $G_1(\theta_1^*) = (1 - \theta_1^*)I + \theta_1^* M_1$. This is shown in Figure 6, together with the spectrum of $G_2(\theta_2^*) = (1 - \theta_2^*)I + \theta_2^* M_2$. We also show in this figure (the right picture) the spectrum of $G_3(\theta_3^*)$, where θ_3^* is the optimal parameter. This parameter will be determined later, in Theorem 4.6.

Lemma 4.4 *Consider $\bar{\theta}_j$, $j = 1, \ldots, p$, as defined in* (31)*. If $\theta \in [\bar{\theta}_l, \bar{\theta}_{l+1}]$, then the eigenvalues $\lambda_j^+(\theta)$ and $\lambda_j^-(\theta)$, $j = 1, \ldots, l$, are real numbers satisfying*

$$\lambda_1^-(\theta) \leq \cdots \leq \lambda_l^-(\theta) \leq \theta - 1 \leq \lambda_l^+(\theta) \leq \cdots \leq \lambda_1^+(\theta) \leq 0.$$

On the other hand, for $j = l + 1, \ldots, p$ we have

$$|\lambda_j^+(\theta)| = |\lambda_j^-(\theta)| = 1 - \theta$$

Thus, the spectral radius of $G_3(\theta)$ is $-\lambda_1^-(\theta)$.

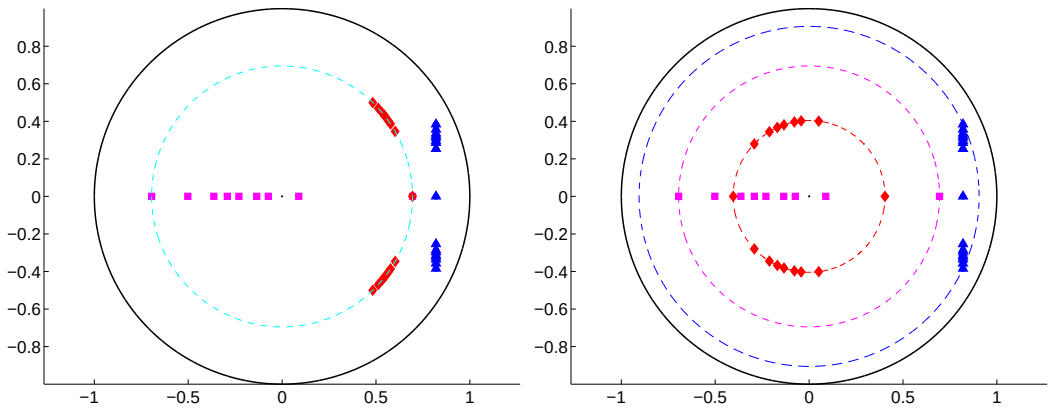

Figure 6: On the left, the spectrum of $G_1(\theta_1^*)$ (magenta), $G_2(\theta_2^*)$ (blue) and $G_3(\theta_1^*)$ (red). On the right, the spectrum of $G_3(\theta_3^*)$, where θ_3^* is the optimal parameter.

From Lemmas 4.3 and 4.4 we can conclude that $\bar{\theta}_1$ is the threshold value for θ after which the eigenvalues of $G_3(\theta)$ start departing the circle of radius $1-\theta$. The next result presents the threshold after which the eigenvalues are all real.

Lemma 4.5 *Consider $\bar{\theta}_p$ as defined in (31). If $\theta \geq \bar{\theta}_p$, then the eigenvalues $\lambda_j^+(\theta)$ and $\lambda_j^-(\theta)$, $j = 1, \ldots, p$, are real numbers satisfying*

$$\lambda_1^-(\theta) \leq \cdots \leq \lambda_p^-(\theta) \leq \theta - 1 \leq \lambda_p^+(\theta) \leq \cdots \leq \lambda_1^+(\theta) \leq 0.$$

Thus, the spectral radius of $G_3(\theta)$ is $-\lambda_1^-(\theta)$.

Proof. The same presented for Lemma 4.4. □

Using the previous results, we can finally establish the convergence of Algorithm 4.1 (Deterministic Quartz).

Theorem 4.6 *Let $w^0 \in \mathbb{R}^d$ and $\alpha^0 \in \mathbb{R}^N$ be arbitrary and consider the sequence $(w^k, \alpha^k)_{k\in\mathbb{N}}$ generated by Algorithm 4.1 with $\theta \in \left(0, \dfrac{2\sqrt{\lambda n}}{\sqrt{\lambda n} + \sigma_1}\right)$. Then the sequence (w^k, α^k) converges to the (unique) solution of the problem (4) at an asymptotic linear rate of*

$$\rho_3(\theta) \overset{\text{def}}{=} \begin{cases} 1 - \theta, \text{ if } \theta \in (0, \bar{\theta}_1] \\ \dfrac{1}{2\lambda n}\left(\theta\sigma_1\sqrt{\delta_1(\theta)} + \theta^2\sigma_1^2 - 2(1-\theta)\lambda n\right), \text{ if } \theta \geq \bar{\theta}_1, \end{cases}$$

where

$$\bar{\theta}_1 = \frac{-2\lambda n + 2\sqrt{\lambda n(\lambda n + \sigma_1^2)}}{\sigma_1^2}.$$

Furthermore, if we choose $\theta_3^* \overset{\text{def}}{=} \bar{\theta}_1$, then the (theoretical) convergence rate is optimal and it is equal to

$$\rho_3^* \overset{\text{def}}{=} 1 - \theta_3^*.$$

It is worth noting that if the spectral radius of M_1 is less than 1, that is, if $\sigma_1^2 < \lambda n$, then Algorithms 3.2, 3.3 and 4.1 converge for any choice of $\theta \in (0, 1]$. Indeed, in this case we have

$$\frac{2\lambda n}{\lambda n + \sigma_1^2} > \frac{2\sqrt{\lambda n}}{\sqrt{\lambda n} + \sigma_1} > 1,$$

which implies that the set of admissible values for θ established in Theorems 3.3, 3.4 and 4.6 contains the whole interval $(0, 1]$.

On the other hand, if $\sigma_1^2 \geq \lambda n$, the convergence of these algorithms is more restrictive. Moreover, in this case we have

$$\frac{2\lambda n}{\lambda n + \sigma_1^2} \leq \frac{2\sqrt{\lambda n}}{\sqrt{\lambda n} + \sigma_1} \leq 1,$$

which means that Algorithm 4.1 has a broader range for θ than Algorithms 3.2 and 3.3.

4.2 Complexity results

Taking into account (22), (25), the relation $\log(1 - \theta) \approx -\theta$ and Theorem 4.6, we conclude that the complexity of our Accelerated Fixed Point Method, Algorithm 4.1, is proportional to

$$\frac{1}{2\theta_3^*} = \frac{\sigma_1^2}{-4\lambda n + 4\sqrt{\lambda n(\lambda n + \sigma_1^2)}} \overset{(25)}{=} \frac{\kappa - \lambda n}{4(\sqrt{\lambda n \kappa} - \lambda n)}. \tag{35}$$

Note that in the case when $\lambda = 1/n$, as is typical in machine learning applications, we can write

$$\frac{1}{2\theta_3^*} \overset{(35)}{=} \frac{\kappa - 1}{4(\sqrt{\kappa} - 1)} = \frac{\sqrt{\kappa} + 1}{4}.$$

This is very surprising as it means that we are achieving the optimal accelerated Nesterov rate $\tilde{O}(\sqrt{\kappa})$.

5 Extensions

In this section we discuss some variants of Algorithm 4.1. The first one consists of switching the order of the computations, updating the dual variable first and then the primal one.

The second approach updates the primal variable enforcing the first relation of the optimality conditions given by (10) and using the relaxation parameter θ only to update the dual variable.

5.1 Switching the update order

This approach updates the dual variable α first and then updates the primal variable w using the new information about α. This is summarized in the following scheme.

$$\begin{cases} \alpha^{k+1} = (1 - \theta)\alpha^k + \theta(y - A^T w^k) \\ w^{k+1} = (1 - \theta)w^k + \theta \dfrac{1}{\lambda n} A\alpha^{k+1}. \end{cases} \tag{36}$$

As we shall see now, this scheme provides the same complexity results as Algorithm 4.1. To see this, note that the iteration (36) is equivalent to

$$\begin{pmatrix} I & -\frac{\theta}{\lambda n}A \\ 0 & I \end{pmatrix}\begin{pmatrix} w^{k+1} \\ \alpha^{k+1} \end{pmatrix} = \begin{pmatrix} (1-\theta)I & 0 \\ -\theta A^T & (1-\theta)I \end{pmatrix}\begin{pmatrix} w^k \\ \alpha^k \end{pmatrix} + \begin{pmatrix} 0 \\ \theta y \end{pmatrix}$$

or in a compact way,

$$x^{k+1} = G(\theta)x^k + f \tag{37}$$

with

$$G(\theta) = \begin{pmatrix} I & -\frac{\theta}{\lambda n}A \\ 0 & I \end{pmatrix}^{-1}\begin{pmatrix} (1-\theta)I & 0 \\ -\theta A^T & (1-\theta)I \end{pmatrix} = (1-\theta)I + \theta\begin{pmatrix} -\frac{\theta}{\lambda n}AA^T & \frac{1-\theta}{\lambda n}A \\ -A^T & 0 \end{pmatrix}. \tag{38}$$

It can be shown that the matrix $G(\theta)$ has exactly the same spectrum of $G_3(\theta)$, defined in (29). So, the convergence result is also the same, which we state again for convenience.

Theorem 5.1 *Let $w^0 \in \mathbb{R}^d$ and $\alpha^0 \in \mathbb{R}^N$ be arbitrary and consider the sequence $(w^k, \alpha^k)_{k\in\mathbb{N}}$ defined by (36) with $\theta \in \left(0, \dfrac{2\sqrt{\lambda n}}{\sqrt{\lambda n}+\sigma_1}\right)$. Then the sequence (w^k, α^k) converges to the (unique) solution of the problem (4) at an asymptotic linear rate of*

$$\rho_3(\theta) = \begin{cases} 1-\theta, \text{ if } \theta \in (0, \bar{\theta}_1] \\ \dfrac{1}{2\lambda n}\left(\theta\sigma_1\sqrt{\delta_1(\theta)} + \theta^2\sigma_1^2 - 2(1-\theta)\lambda n\right), \text{ if } \theta \geq \bar{\theta}_1, \end{cases}$$

where

$$\bar{\theta}_1 = \frac{-2\lambda n + 2\sqrt{\lambda n(\lambda n + \sigma_1^2)}}{\sigma_1^2}.$$

Furthermore, if we choose $\theta_3^* = \bar{\theta}_1$, then the (theoretical) convergence rate is optimal and it is equal to

$$\rho_3^* = 1 - \theta_3^*.$$

5.2 Maintaining primal-dual relationship

The second approach updates the primal variable enforcing the first relation of the optimality conditions given by (10) and uses the relaxation parameter θ only to update the dual variable, as described in the following scheme.

$$\begin{cases} w^{k+1} = \dfrac{1}{\lambda n}A\alpha^k \\ \alpha^{k+1} = (1-\theta)\alpha^k + \theta(y - A^T w^{k+1}). \end{cases} \tag{39}$$

Differently from the previous case, this scheme cannot achieve accelerated convergence.

As we did before, note that the scheme (39) can be written as

$$\begin{pmatrix} I & 0 \\ \theta A^T & I \end{pmatrix}\begin{pmatrix} w^{k+1} \\ \alpha^{k+1} \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{\lambda n}A \\ 0 & (1-\theta)I \end{pmatrix}\begin{pmatrix} w^k \\ \alpha^k \end{pmatrix} + \begin{pmatrix} 0 \\ \theta y \end{pmatrix}$$

or in a compact way,

$$x^{k+1} = G(\theta)x^k + f \tag{40}$$

with

$$G(\theta) = \begin{pmatrix} I & 0 \\ \theta A^T & I \end{pmatrix}^{-1} \begin{pmatrix} 0 & \frac{1}{\lambda n}A \\ 0 & (1-\theta)I \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{\lambda n}A \\ 0 & (1-\theta)I - \frac{\theta}{\lambda n}A^T A \end{pmatrix}. \tag{41}$$

We can conclude that the eigenvalues of this matrix are

$$\left\{1 - \theta - \frac{\theta\sigma_j^2}{\lambda n}\ ,\ j = 1, \ldots, p\right\} \cup \{1-\theta\},$$

exactly the same of the matrix $G_1(\theta)$, the iteration matrix of Algorithm 3.1 with employment of M_1. So, the complexity analysis here is the same as that one established in Theorem 3.3.

5.3 Maintaining primal-dual relationship 2

For the sake of completeness, we present next the method where we keep the second relationship intact and include θ in the first relationship. This leads to

$$\begin{cases} \alpha^{k+1} = y - A^T w^k \\ w^{k+1} = (1-\theta)w^k + \dfrac{\theta}{\lambda n}A\alpha^{k+1}. \end{cases} \tag{42}$$

Here we obtain the same convergence results as the ones described in Section 5.2. In fact, the relations above can be written as

$$\begin{pmatrix} 0 & I \\ I & -\frac{\theta}{\lambda n}A \end{pmatrix} \begin{pmatrix} w^{k+1} \\ \alpha^{k+1} \end{pmatrix} = \begin{pmatrix} -A^T & 0 \\ (1-\theta)I & 0 \end{pmatrix} \begin{pmatrix} w^k \\ \alpha^k \end{pmatrix} + \begin{pmatrix} y \\ 0 \end{pmatrix}$$

or in a compact way,

$$x^{k+1} = G(\theta)x^k + f \tag{43}$$

with

$$G(\theta) = \begin{pmatrix} 0 & I \\ I & -\frac{\theta}{\lambda n}A \end{pmatrix}^{-1} \begin{pmatrix} -A^T & 0 \\ (1-\theta)I & 0 \end{pmatrix} = \begin{pmatrix} (1-\theta)I - \frac{\theta}{\lambda n}AA^T & 0 \\ -A^T & 0 \end{pmatrix}. \tag{44}$$

We can conclude that the eigenvalues of this matrix are

$$\left\{1 - \theta - \frac{\theta\sigma_j^2}{\lambda n}\ ,\ j = 1, \ldots, p\right\} \cup \{1-\theta\},$$

exactly the same of the matrix $G_1(\theta)$, the iteration matrix of Algorithm 3.1 with employment of M_1. So, the complexity analysis here is the same as that one established in Theorem 3.3.

Observe that in (39) we have

$$w^{k+1} = \phi_1(\alpha^k) \quad \text{and} \quad \alpha^{k+1} = \phi_2(\theta, \alpha^k, w^{k+1}).$$

On the other hand, in (42) we have

$$\alpha^{k+1} = \phi_3(w^k) \quad \text{and} \quad w^{k+1} = \phi_4(\theta, w^k, \alpha^{k+1}).$$

It is worth noting that if we update the variables as

$$\alpha^{k+1} = \phi_2(\theta, \alpha^k, w^k) \quad \text{and} \quad w^{k+1} = \phi_1(\alpha^{k+1})$$

or

$$w^{k+1} = \phi_4(\theta, w^k, \alpha^k) \quad \text{and} \quad \alpha^{k+1} = \phi_3(w^{k+1})$$

we obtain

$$\begin{pmatrix} -\frac{\theta}{\lambda n} AA^T & \frac{(1-\theta)}{\lambda n} A \\ -\theta A^T & (1-\theta) I \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} (1-\theta) I & \frac{\theta}{\lambda n} A \\ -(1-\theta) A^T & -\frac{\theta}{\lambda n} A^T A \end{pmatrix}$$

as the associated iteration matrices, respectively. Moreover, we can conclude that they also have the same spectrum of $G_1(\theta)$. So, the complexity analysis is the same as that one established in Theorem 3.3.

6 Numerical Experiments

In this section we present a comparison among the methods discussed in this work. Besides a table with the convergence rates and complexity bounds, we show here some numerical tests performed to illustrate the properties of Algorithms 3.2, 3.2 and 4.1 as well as of the extensions (36) and (39) applied to solve the primal-dual ridge regression problem stated in (4). We refer to Algorithm 4.1 as Quartz and the extensions (36) and (39) as New Quartz and Modified Quartz, respectively.

We summarize the main features of these methods in Table 2 which brings the range for the parameter to ensure convergence, the optimal convergence rates, the complexity and the cost per iteration of each method. For instance, the two versions of Algorithm 3.1 have the same range for theta. The usage of M_1 provides best convergence rate compared with using M_2. However, it requires more calculations per iteration: the major computational tasks to be performed are computation of the matrix-vector products $AA^T w$ and $A^T A\alpha$, while the use of M_2 needs the computation of $A\alpha$ and $A^T w$.

Surprisingly, Algorithm 4.1 has shown to be the best from both the theoretical point of view and the numerical experiments and with the same cost as the computation of $A\alpha$ and $A^T w$.

We also point out that the modified Quartz, (39), did not have here the same performance as the randomized version studied in [8].

Figures 7 and 8 illustrate these features, showing the primal-dual objective values against the number of iterations. The dimensions considered were $d = 10$, $m = 1$ and $n = 500$. We adopted the optimal parameters associated with each method. Figure 7 compares the two variants of Algorithm 3.1 against Algorithm 4.1, while Figure 8 shows the three variants of algorithm Quartz and MFP(M_1). We can see the equivalence between Quartz and New Quartz and also the equivalence between Modified Quartz and MFP(M_1).

	MFP(M_1, θ)	MFP(M_2, θ)	Qtz(θ)	NewQtz(θ)	ModQtz(θ)
Range for θ	$\left(0, \frac{2\lambda n}{\lambda n + \sigma_1^2}\right)$	$\left(0, \frac{2\lambda n}{\lambda n + \sigma_1^2}\right)$	$\left(0, \frac{2\sqrt{\lambda n}}{\sqrt{\lambda n} + \sigma_1}\right)$	$\left(0, \frac{2\sqrt{\lambda n}}{\sqrt{\lambda n} + \sigma_1}\right)$	$\left(0, \frac{2\lambda n}{\lambda n + \sigma_1^2}\right)$
Optimal rate	$\frac{\sigma_1^2}{2\lambda n + \sigma_1^2}$	$\sqrt{\frac{\sigma_1^2}{\lambda n + \sigma_1^2}}$	$1 - \theta_3^*$	$1 - \theta_3^*$	$\frac{\sigma_1^2}{2\lambda n + \sigma_1^2}$
Complexity	(26)	(27)	(35)	(35)	(26)
Cost/iteration	AA^Tw, $A^TA\alpha$	$A\alpha$, A^Tw	$A\alpha$, A^Tw^+	A^Tw, $A\alpha^+$	$A\alpha$, A^Tw^+

Table 2: Comparison between the convergence rates and cost per iteration of the different versions of Algorithms 3.1 and 4.1.

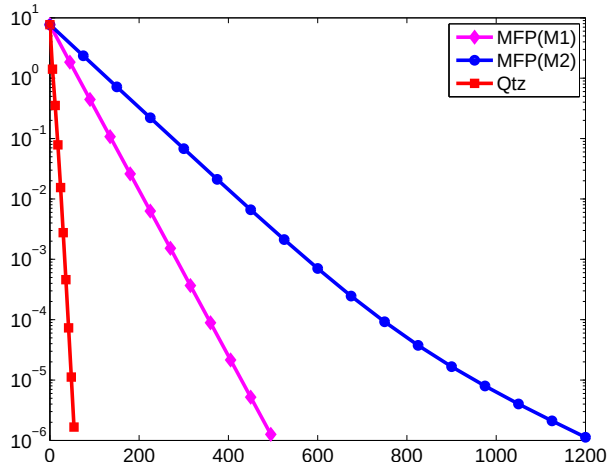

Figure 7: Comparison among the two versions of Algorithm 3.1 and Deterministic Quartz.

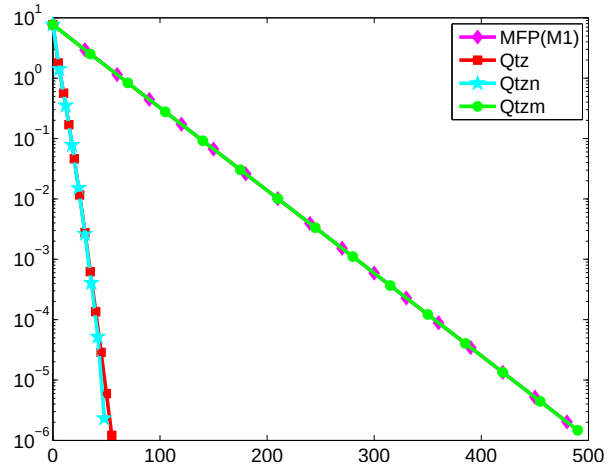

Figure 8: Comparison among the different versions of Deterministic Quartz and Algorithm 3.2.

7 Conclusion

In this paper we have proposed and analyzed several algorithms for solving the ridge regression problem. The proposed methods are based on the optimality conditions for the problem of minimizing the duality gap between the primal and dual ridge regression problems. In fact, we have described a family of (parameterized) fixed point methods applied to the reformulations for the optimality conditions. We have presented the convergence analysis and complexity results for these methods and proved that our main method achieves the optimal accelerated Nesterov rate. We have performed some numerical experiments to illustrate the properties of our algorithms.

References

[1] K J Arrow and L Hurwicz. Gradient method for concave programming I: Local results. In K J Arrow, L Hurwicz, and H Uzawa, editors, *Studies in Linear and Nonlinear Programming*, pages 117–126. Stanford University Press, Stanford, 1958.

[2] A Chambolle and T Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40:120–145, 2011.

[3] M El-Dereny and N I Rashwan. Solving multicollinearity problem using ridge regression models. *Int. J. Contemp. Math. Sci.*, 6:585–600, 2011.

[4] D M Hawkins and X Yin. A faster algorithm for ridge regression of reduced rank data. *Comput. Statist. Data Anal.*, 40(2):253–262, 2002.

[5] A E Hoerl. Application of ridge analysis to regression problems. *Chem. Eng. Prog.*, 58:54–59, 1962.

[6] A E Hoerl and R W Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[7] N Komodakis and J C Pesquet. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Process. Mag.*, 32(6):31–54, 2015.

[8] Z Qu, P Richtárik, and T Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *Adv. Neural Inf. Process. Syst. 28*, pages 865–873, 2015.

[9] C Saunders, A Gammerman, and V Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann Publishers Inc., 1998.

[10] S Shalev-Shwartz and T Zhang. Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 14(1):567–599, 2013.

[11] S Shalev-Shwartz and T Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Math. Program., Ser. A*, 155:105–145, 2016.

[12] T C Silva, A A Ribeiro, and G A Periçaro. A new accelerated algorithm for ill-conditioned ridge regression problems. *Comp. Appl. Math.*, 2017. To appear.

[13] H Uzawa. Iterative methods for concave programming. In K J Arrow, L Hurwicz, and H Uzawa, editors, *Studies in Linear and Nonlinear Programming*, pages 117–126. Stanford University Press, Stanford, 1958.

[14] H D Vinod. A survey of ridge regression and related techniques for improvements over ordinary least squares. *The Review of Economics and Statistics*, 60(1):121–131, 1978.

[15] T Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46(1):91–129, 2002.

A Proofs of Results from Section 3

First, let us see a basic linear algebra result whose proof is straightforward by induction.

Proposition A.1 *Let $Q_j \in \mathbb{R}^{l\times l}$, $j = 1,\ldots,4$, be diagonal matrices whose diagonal entries are components of $\alpha,\beta,\gamma,\delta \in \mathbb{R}^l$, respectively. Then*

$$\det\begin{pmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{pmatrix} = \prod_{j=1}^{l} (\alpha_j\delta_j - \beta_j\gamma_j)\,.$$

A.1 Proof of Lemma 3.1

Let $c = -\dfrac{1}{\lambda n}$. From (18) and (19), we can write $M_1 = W\Sigma_1 W^T$ and $M_2 = W\Sigma_2 W^T$, where

$$W = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix},\ \Sigma_1 = \begin{pmatrix} c\Sigma\Sigma^T & 0 \\ 0 & c\Sigma^T\Sigma \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{array}{c} \\ \begin{pmatrix} 0 & 0 & -c\widetilde{\Sigma} & 0 \\ 0 & 0 & 0 & 0 \\ -\widetilde{\Sigma} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ \begin{matrix} p & d-p & p & N-p \end{matrix} \end{array} \begin{array}{c} \\ p \\ d-p \\ p \\ N-p \end{array}$$

The evaluation of $p_1(t) = \det(tI - M_1) = \det(tI - \Sigma_1)$ is straightforward and

$$p_2(t) = \det(tI - M_2) = \det\begin{pmatrix} tI & 0 & c\widetilde{\Sigma} & 0 \\ 0 & tI & 0 & 0 \\ \widetilde{\Sigma} & 0 & tI & 0 \\ 0 & 0 & 0 & tI \end{pmatrix} = \det\begin{pmatrix} tI & c\widetilde{\Sigma} & 0 & 0 \\ \widetilde{\Sigma} & tI & 0 & 0 \\ 0 & 0 & tI & 0 \\ 0 & 0 & 0 & tI \end{pmatrix}.$$

The result then follows from Proposition A.1.

A.2 Proof of Theorem 3.3

We claim that the spectral radius of $G_1(\theta) \stackrel{\text{def}}{=} (1-\theta)I + \theta M_1$ is $\rho_1(\theta)$ and also coincides with $\|G_1(\theta)\|$. Using Lemma 3.1, we conclude that the eigenvalues of this matrix are

$$\left\{1 - \theta - \frac{\theta\sigma_j^2}{\lambda n}\,, \quad j = 1,\ldots,p\right\} \cup \{1-\theta\}.$$

So, its spectral radius is

$$\max\left\{\left|1 - \theta\left(1 + \frac{\sigma_1^2}{\lambda n}\right)\right|, 1-\theta\right\} = \rho_1(\theta).$$

Since $G_1(\theta)$ is symmetric, this quantity coincides with $\|G_1(\theta)\|$. Furthermore, the admissible values for θ, that is, the ones such that the eigenvalues have modulus less than one, can be found by solving

$$\left|1 - \theta\left(1 + \frac{\sigma_1^2}{\lambda n}\right)\right| < 1,$$

which immediately gives $0 < \theta < \dfrac{2\lambda n}{\lambda n + \sigma_1^2}$. So, the linear convergence of Algorithm 3.1 is guaranteed for any $\theta \in \left(0, \dfrac{2\lambda n}{\lambda n + \sigma_1^2}\right)$. Finally, note that the solution of

$$\min_{\theta>0} \rho_1(\theta)$$

is achieved when

$$\theta\left(1 + \frac{\sigma_1^2}{\lambda n}\right) - 1 = 1 - \theta,$$

yielding $\theta_1^* = \dfrac{2\lambda n}{2\lambda n + \sigma_1^2}$ and the optimal convergence rate $\rho_1^* = \dfrac{\sigma_1^2}{2\lambda n + \sigma_1^2}$.

A.3 Proof of Theorem 3.4

First, using Lemma 3.1, we conclude that the eigenvalues of $G_2(\theta) \overset{\text{def}}{=} (1-\theta)I + \theta M_2$ are

$$\left\{1 - \theta \pm \frac{\theta\sigma_j}{\sqrt{\lambda n}} i\,, \quad j = 1, \ldots, p\right\} \cup \{1 - \theta\},$$

where $i = \sqrt{-1}$. The two ones with largest modulus are $1 - \theta \pm \dfrac{\theta\sigma_1}{\sqrt{\lambda n}} i$ (see Figure 2). So, the spectral radius of $G_2(\theta)$ is

$$\sqrt{(1-\theta)^2 + \frac{\theta^2\sigma_1^2}{\lambda n}} = \rho_2(\theta).$$

Further, the values of θ for which the eigenvalues of $G_2(\theta)$ have modulus less than one can be found by solving $(1-\theta)^2 + \dfrac{\theta^2\sigma_1^2}{\lambda n} < 1$ giving

$$0 < \theta < \frac{2\lambda n}{\lambda n + \sigma_1^2}.$$

The asymptotic convergence follows from the fact that $\|G_2(\theta)^k\|^{1/k} \to \rho_2(\theta)$. Indeed, using (17) we conclude that

$$\left(\frac{\|x^k - x^*\|}{\|x^0 - x^*\|}\right)^{1/k} \leq \|G_2(\theta)^k\|^{1/k} \to \rho_2(\theta).$$

This means that given $\gamma > 0$, there exists $k_0 \in \mathbb{N}$ such that

$$\|x^k - x^*\| \leq (\rho_2(\theta) + \gamma)^k \|x^0 - x^*\|$$

for all $k \geq k_0$. Finally, the optimal parameter θ_2^* and the corresponding optimal rate ρ_2^* can be obtained directly by solving

$$\min_{\theta>0} (1-\theta)^2 + \frac{\theta^2\sigma_1^2}{\lambda n}.$$

B Proofs of Results from Section 4

B.1 Proof of Lemma 4.1

Consider the matrix

$$M_3(\theta) \stackrel{\text{def}}{=} \begin{pmatrix} 0 & \frac{1}{\lambda n} A \\ (\theta-1)A^T & -\frac{\theta}{\lambda n} A^T A \end{pmatrix}.$$

Using the singular value decomposition of A, given in (18), we can write

$$M_3(\theta) = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{\lambda n}\Sigma \\ (\theta-1)\Sigma^T & -\frac{\theta}{\lambda n}\Sigma^T\Sigma \end{pmatrix} \begin{pmatrix} U^T & 0 \\ 0 & V^T \end{pmatrix}.$$

Therefore, the eigenvalues of $M_3(\theta)$ are the same as the ones of

$$\begin{pmatrix} 0 & \frac{1}{\lambda n}\Sigma \\ (\theta-1)\Sigma^T & -\frac{\theta}{\lambda n}\Sigma^T\Sigma \end{pmatrix} = \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} \begin{pmatrix} 0 & 0 & -c\widetilde{\Sigma} & 0 \\ 0 & 0 & 0 & 0 \\ (\theta-1)\widetilde{\Sigma} & 0 & \theta c\widetilde{\Sigma}^2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ \begin{matrix} p & d-p & p & N-p \end{matrix} \end{matrix} \begin{matrix} p \\ d-p \\ p \\ N-p \end{matrix}$$

where $c = -\dfrac{1}{\lambda n}$ and $\widetilde{\Sigma}$ is defined in (19). The characteristic polynomial of this matrix is

$$p_\theta(t) = \det \begin{pmatrix} tI & 0 & c\widetilde{\Sigma} & 0 \\ 0 & tI & 0 & 0 \\ (1-\theta)\widetilde{\Sigma} & 0 & tI - \theta c\widetilde{\Sigma}^2 & 0 \\ 0 & 0 & 0 & tI \end{pmatrix} = \det \begin{pmatrix} tI & c\widetilde{\Sigma} & 0 & 0 \\ (1-\theta)\widetilde{\Sigma} & tI - \theta c\widetilde{\Sigma}^2 & 0 & 0 \\ 0 & 0 & tI & 0 \\ 0 & 0 & 0 & tI \end{pmatrix}.$$

Using Proposition A.1, we obtain

$$p_\theta(t) = t^{N+d-2p} \prod_{j=1}^{p} \Big(t(t - \theta c\sigma_j^2) - c(1-\theta)\sigma_j^2 \Big) = t^{N+d-2p} \prod_{j=1}^{p} \left(t^2 + \frac{\theta\sigma_j^2}{\lambda n} t + \frac{(1-\theta)\sigma_j^2}{\lambda n} \right).$$

Thus, the eigenvalues of $M_3(\theta)$ are

$$\frac{1}{2\lambda n} \left\{ -\theta\sigma_j^2 \pm \sigma_j \sqrt{\theta^2\sigma_j^2 - 4(1-\theta)\lambda n}\,, \quad j = 1, \ldots, p \right\} \cup \{0\},$$

so that the eigenvalues of $G_3(\theta) = (1-\theta)I + \theta M_3(\theta)$ are

$$\frac{1}{2\lambda n} \left\{ 2(1-\theta)\lambda n - \theta^2\sigma_j^2 \pm \theta\sigma_j \sqrt{\theta^2\sigma_j^2 - 4(1-\theta)\lambda n}\,, \quad j = 1, \ldots, p \right\} \cup \{1-\theta\},$$

giving the desired result.

B.2 Proof of Lemma 4.3

Note that in this case we have $\delta_j(\theta) \leq 0$ for all $j = 1, \ldots, p$. So,

$$\begin{aligned}
|\lambda_j^+(\theta)|^2 &= |\lambda_j^-(\theta)|^2 \\
&= \frac{1}{4\lambda^2 n^2}\left(\left(2(1-\theta)\lambda n - \theta^2\sigma_j^2\right)^2 - \theta^2\sigma_j^2\delta_j(\theta)\right) \\
&= \frac{1}{4\lambda^2 n^2}\left(4(1-\theta)^2\lambda^2 n^2 - 4(1-\theta)\lambda n\theta^2\sigma_j^2 + \theta^4\sigma_j^4 - \theta^2\sigma_j^2\big(\theta^2\sigma_j^2 - 4(1-\theta)\lambda n\big)\right) \\
&= (1-\theta)^2,
\end{aligned}$$

yielding the desired result since $\theta \leq 1$.

B.3 Proof of Lemma 4.4

We have $\delta_j(\theta) \geq 0$ for all $j = 1, \ldots, l$. So,

$$\begin{aligned}
\lambda_j^+(\theta) - (\theta - 1) &= \frac{1}{2\lambda n}\left(2(1-\theta)\lambda n - \theta^2\sigma_j^2 + \theta\sigma_j\sqrt{\delta_j(\theta)}\right) + 1 - \theta \\
&= \frac{1}{2\lambda n}\left(4(1-\theta)\lambda n - \theta^2\sigma_j^2 + \theta\sigma_j\sqrt{\delta_j(\theta)}\right) \\
&= \frac{1}{2\lambda n}\left(-\delta_j(\theta) + \theta\sigma_j\sqrt{\delta_j(\theta)}\right) \\
&= \frac{\sqrt{\delta_j(\theta)}}{2\lambda n}\left(\theta\sigma_j - \sqrt{\delta_j(\theta)}\right) \geq 0.
\end{aligned}$$

Furthermore,

$$\left(\theta\sigma_j\sqrt{\delta_j(\theta)}\right)^2 = \theta^2\sigma_j^2\left(\theta^2\sigma_j^2 - 4(1-\theta)\lambda n\right) \leq \left(\theta^2\sigma_j^2 - 2(1-\theta)\lambda n\right)^2.$$

Since $\theta^2\sigma_j^2 - 2(1-\theta)\lambda n = \delta_j(\theta) + 2(1-\theta)\lambda n \geq 0$,

$$\lambda_j^+(\theta) = \frac{1}{2\lambda n}\left(2(1-\theta)\lambda n - \theta^2\sigma_j^2 + \theta\sigma_j\sqrt{\delta_j(\theta)}\right) \leq 0.$$

Now, note that

$$\begin{aligned}
\lambda_j^-(\theta) - (\theta - 1) &= \frac{1}{2\lambda n}\left(2(1-\theta)\lambda n - \theta^2\sigma_j^2 - \theta\sigma_j\sqrt{\delta_j(\theta)}\right) + 1 - \theta \\
&= \frac{1}{2\lambda n}\Big(-\delta_j(\theta) - \theta\sigma_j\sqrt{\delta_j(\theta)}\Big) \leq 0.
\end{aligned}$$

Moreover, from Lemma 4.2 and the definition of σ_j, we have $\delta_1(\theta) \geq \cdots \geq \delta_l(\theta)$ and $\sigma_1 \geq \cdots \geq \sigma_l$, which imply that $\lambda_1^-(\theta) \leq \cdots \leq \lambda_l^-(\theta)$. The inequality $\lambda_l^+(\theta) \leq \cdots \leq \lambda_1^+(\theta)$ follows from the fact that the function

$$[\sqrt{a}, \infty) \ni s \mapsto -s^2 + s\sqrt{s^2 - a}$$

is increasing. Finally, for $j = l+1, \ldots, p$ we have $\delta_j(\theta) \leq 0$ and, by the same argument used in Lemma 4.3, we conclude that $|\lambda_j^+(\theta)| = |\lambda_j^-(\theta)| = 1 - \theta$.

B.4 Proof of Theorem 4.6

Since Algorithm 4.1 can be represented by (28), we need to show that $\rho(G_3(\theta))$, the spectral radius of $G_3(\theta)$, is less than 1. First, note that by Lemmas 4.3, 4.4 and 4.5, we have $\rho(G_3(\theta)) = \rho_3(\theta)$. Using Lemma 4.2 we conclude that the function $\theta \mapsto \rho_3(\theta)$ is increasing on the interval $[\bar{\theta}_1, \infty)$, which means that its minimum is attained at $\bar{\theta}_1$. To finish the proof, it is enough to prove that $\rho_3(\theta) = 1$ if and only if

$$\theta = \frac{2\sqrt{\lambda n}}{\sqrt{\lambda n} + \sigma_1}.$$

Note that

$$\begin{aligned}
\rho_3(\theta) = 1 \quad &\Leftrightarrow \quad \theta\sigma_1\sqrt{\delta_1(\theta)} + \theta^2\sigma_1^2 - 2(1-\theta)\lambda n = 2\lambda n \\
&\Rightarrow \quad \theta^2\sigma_1^2\delta_1(\theta) = \left(2(2-\theta)\lambda n - \theta^2\sigma_1^2\right)^2 \\
&\Leftrightarrow \quad \frac{2-\theta}{\theta} = \frac{\sigma_1}{\sqrt{\lambda n}} \\
&\Leftrightarrow \quad \theta = \frac{2\sqrt{\lambda n}}{\sqrt{\lambda n} + \sigma_1}
\end{aligned}$$

and

$$\begin{aligned}
\theta = \frac{2\sqrt{\lambda n}}{\sqrt{\lambda n} + \sigma_1} \quad &\Leftrightarrow \quad \theta^2\sigma_1^2 = \lambda n(2-\theta)^2 \\
&\Leftrightarrow \quad \theta\sigma_1\sqrt{\delta_1(\theta)} + \theta^2\sigma_1^2 - 2(1-\theta)\lambda n = 2\lambda n,
\end{aligned}$$

completing the proof.