Block Acceleration Without Momentum: On Optimal Stepsizes of Block Gradient Descent for Least-Squares

Liangzu Peng [1 *] **Wotao Yin** [2]

Abstract

Block coordinate descent is a powerful algorithmic template suitable for big data optimization. This template admits a lot of variants including block gradient descent (BGD), which performs gradient descent on a selected block of variables, while keeping other variables fixed. For a very long time, the stepsize for each block has tacitly been set to one divided by the block-wise Lipschitz smoothness constant, imitating the vanilla stepsize rule for gradient descent (GD). However, such a choice for BGD has not yet been able to theoretically justify its empirical superiority over GD, as existing convergence rates for BGD have worse constants than GD in the deterministic cases.

To discover such theoretical justification, we set up a simple environment where we consider BGD applied to least-squares with two blocks of variables. Assuming the data matrix corresponding to each block is orthogonal, we find optimal stepsizes of BGD in closed form, which provably lead to asymptotic convergence rates twice as fast as GD with Polyak's momentum; this means, under that orthogonality assumption, one can accelerate BGD by just tuning stepsizes and without adding any momentum. An application that satisfies this assumption is *generalized alternating projection* between two subspaces, and applying our stepsizes to it improves the prior convergence rate that was once claimed, slightly inaccurately, to be optimal. The main proof idea is to minimize, in stepsize variables, the spectral radius of a matrix that controls convergence rates.

[*]Work done during a summer internship at Alibaba US DAMO Academy. [1]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA [2]Decision Intelligence Lab, DAMO Academy, Alibaba Group US, Bellevue, WA 98004, USA. Correspondence to: Liangzu Peng <lpenn@seas.upenn.edu>, Wotao Yin <wotao.yin@alibaba-inc.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Block coordinate descent refers to a family of algorithms selecting and updating one block of variables at a time. In the span of more than six decades since its early appearance (Hildreth, 1957), many variants of block coordinate descent have been proposed, analyzed, and recently tested on big data scenarios (Nesterov, 2012; Xu & Yin, 2013; Wright, 2015; Shi et al., 2016; Lin et al., 2023). Despite the abundance of these exciting developments, one might ponder when one should prefer block coordinate descent to vanilla gradient descent (GD). Intuitively, block coordinate descent might be advantageous if the optimization variables admit a natural partition into blocks, which is the case in many applications (Peng & Vidal, 2023); or if the variable dimension is too high to fit the memory, in which case one is forced to optimize in a block-wise manner (Nesterov, 2012); or if the given optimization problem is *coordinate-friendly* (Peng et al., 2016), meaning that minimizing over one block of coordinates while fixing others is computationally easy.

Developing convergence theory to support the above intuition, however, is challenging. For example, consider *block gradient descent* (BGD), a method that runs GD at every iteration on a selected block of variables. Nesterov was concerned that, if selecting blocks in a *greedy* fashion (e.g., using the famous *Gauss–Southwell* rule), then, following standard reasoning, one obtains a bound on convergence rates that might have a worse constant than that of GD (Nesterov, 2012). This concern led him to a BGD variant that randomly selects blocks, whose convergence rate is proved to be better than GD. While theoretically appealing and having attracted a sequence of follow-up works, such a randomized variant ensures convergence only in expectation; this is perhaps why it is empirically slower[1] for the least-squares problem than *cyclic BGD*, a BGD variant that selects blocks in a cyclic fashion, see Tables 3.1 & 3.2 of Beck & Tetruashvili (2013). We focus on the cyclic rule, and for short, we write BGD for cyclic BGD in the sequel.

In the deterministic setting, the *BGD versus GD* dilemma persists: BGD is empirically faster, but its current bound on

[1]It is also shown that this randomized variant is in many cases empirically slower than the greedy version (Nutini et al., 2015).

(deterministic) convergence rates has, in general, a worse constant than GD; see Remark 3.3 of Beck & Tetruashvili (2013). Furthermore, this bittersweet dilemma manifests itself again if one extends, in a direct way, BGD and GD into their proximal versions (e.g., compare Theorems 10.15 and 11.18 of Beck (2017)), or into their accelerated versions (e.g., compare Theorem 4.2 of Beck & Tetruashvili (2013) and Theorem 2.2.2 of Nesterov (2018a)).

How does one even reconcile these? It is now the case that there is some sub-optimality in the existing analysis of BGD methods, but it might also be the case that, after decades of development, making improvements is difficult.

Our approach to making progress features a return to the very basic setting, where we minimize the arguably simplest, very well-studied objective, *least-squares*,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} F(\boldsymbol{x}), \quad F(\boldsymbol{x}) := \frac{1}{2}\cdot\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|_2^2. \tag{1}$$

Here we assume the $m\times n$ matrix A is of full column rank n (necessarily, $m\geq n$). Instead of with multiple, we consider BGD with only two blocks. Specifically, we write $\boldsymbol{x}:=[\boldsymbol{x}_1;\boldsymbol{x}_2]$, $\boldsymbol{A}:=[\boldsymbol{A}_1\ \boldsymbol{A}_2]$, and write F and (1) into

$$\min_{\boldsymbol{x}_1\in\mathbb{R}^{n_1},\boldsymbol{x}_2\in\mathbb{R}^{n_2}} \frac{1}{2}\cdot\|\boldsymbol{A}_1\boldsymbol{x}_1+\boldsymbol{A}_2\boldsymbol{x}_2-\boldsymbol{y}\|_2^2. \tag{2}$$

We run GD on each block in an alternating fashion with constant stepsizes γ_1 and γ_2 to minimize the least-squares objective F and search for the global minimizer $\boldsymbol{x}^*$:

$$\begin{aligned}\boldsymbol{x}_1^+ &= \boldsymbol{x}_1-\gamma_1\cdot\nabla_{\boldsymbol{x}_1}F(\boldsymbol{x})\\ &= \boldsymbol{x}_1-\gamma_1\cdot\boldsymbol{A}_1^\top(\boldsymbol{A}_1\boldsymbol{x}_1+\boldsymbol{A}_2\boldsymbol{x}_2-\boldsymbol{y}),\\ \boldsymbol{x}_2^+ &= \boldsymbol{x}_2-\gamma_2\cdot\nabla_{\boldsymbol{x}_2}F([\boldsymbol{x}_1^+;\boldsymbol{x}_2])\\ &= \boldsymbol{x}_2-\gamma_2\cdot\boldsymbol{A}_2^\top(\boldsymbol{A}_1\boldsymbol{x}_1^+ +\boldsymbol{A}_2\boldsymbol{x}_2-\boldsymbol{y}).\end{aligned} \tag{BGD}$$

We denote by $\boldsymbol{x}^+:=[\boldsymbol{x}_1^+;\boldsymbol{x}_2^+]$ and $\boldsymbol{x}:=[\boldsymbol{x}_1;\boldsymbol{x}_2]$ the two consecutive iterates of BGD. It is clear that the convergence (rate) of BGD depends on stepsizes γ_1,γ_2 and that the fastest convergence of BGD is attained only if we set the stepsizes to be "optimal". Let the word "optimal" be vague for the moment, and we report the following result:

Theorem 1 (Informal)**.** *Suppose Assumption 1 below holds. Run GD with Polyak's momentum (i.e., the heavy ball method) and BGD, respectively, with their "optimal" stepsizes. BGD is twice as fast as the heavy ball method (HB).*

Assumption 1. $\boldsymbol{A}_1^\top\boldsymbol{A}_1$ and $\boldsymbol{A}_2^\top\boldsymbol{A}_2$ are identity matrices, and $\boldsymbol{A}_2^\top\boldsymbol{A}_1\neq 0$.

The rest of the paper is organized as follows. In Section 2 we establish notations and quantify optimal stepsizes. In Section 3 we elaborate on Assumption 1 and argue that Assumption 1 is valid and reasonable. In Section 4 we present optimal stepsizes for BGD in comparison to the heavy ball method, from which Theorem 1 follows. In Section 5 we make connections to prior works, and in particular in Section 5.5 we connect BGD to alternating projection. Then, in Section 6, we apply our results to the problem of alternating projection, resulting in improvements over prior works. In Section 7, we perform basic experiments to verify our theory, and in Section 8 we conclude the paper.

2. Quantifying Optimality

In Section 2.1 we review gradient descent (GD) and the heavy ball method (HB), and clarify what "optimal" stepsizes mean for them. In Section 2.2 we characterize optimal stepsizes for BGD and state Theorem 1 formally.

2.1. Gradient Descent and Heavy Ball: A Review

Recall that vanilla gradient descent applied to least-squares (1) comes with a stepsize $\gamma>0$ and updates $\boldsymbol{x}$ via

$$\boldsymbol{x}^+=\boldsymbol{x}-\gamma\cdot\nabla F(\boldsymbol{x})=\boldsymbol{x}-\gamma\cdot\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}). \tag{GD}$$

The two consecutive iterates $\boldsymbol{x}^+$ and $\boldsymbol{x}$ of GD satisfy

$$\boldsymbol{x}^+-\boldsymbol{x}^*=(\boldsymbol{I}-\gamma\boldsymbol{A}^\top\boldsymbol{A})\cdot(\boldsymbol{x}-\boldsymbol{x}^*),$$

so the rate of convergence to $\boldsymbol{x}^*$ is dictated by the spectrum of $\boldsymbol{I}-\gamma\boldsymbol{A}^\top\boldsymbol{A}$. Thus we wish to find a stepsize γ to minimize the *spectral radius* $\rho(\boldsymbol{I}-\gamma\boldsymbol{A}^\top\boldsymbol{A})$, where $\rho(\cdot)$ denotes the maximum magnitude of eigenvalues of a matrix. Explicitly:

$$\rho_{\text{GD}}(\gamma):=\rho(\boldsymbol{I}-\gamma\boldsymbol{A}^\top\boldsymbol{A})=\max_{i=1,\dots,n}|\lambda_i(\boldsymbol{I}-\gamma\boldsymbol{A}^\top\boldsymbol{A})|.$$

Here we used $\lambda_i(\cdot)$ to mean the i-th largest eigenvalue of a matrix. A folklore fact in optimization is that $\rho_{\text{GD}}(\gamma)$ is minimized at $\gamma=\gamma^*:=\frac{2}{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})+\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})}$, that is

$$\begin{aligned}\rho_{\text{GD}}^*:=\rho_{\text{GD}}(\gamma^*)&=\min_{\gamma>0}\rho_{\text{GD}}(\gamma)\\ &=\frac{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})-\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})}{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})+\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})}.\end{aligned}$$

We say γ^* is the optimal (constant) stepsize of GD applied to least-squares as it minimizes the spectral radius $\rho_{\text{GD}}(\gamma)$.

In a similar style we review the heavy ball method (Polyak, 1964). This method calculates the current iterate $\boldsymbol{x}^{++}$ using two previous points $\boldsymbol{x}^+$ and $\boldsymbol{x}$, that is

$$\boldsymbol{x}^{++}=\boldsymbol{x}^+-\alpha\cdot\nabla F(\boldsymbol{x}^+)+\beta\cdot(\boldsymbol{x}^+-\boldsymbol{x}), \tag{HB}$$

where $\nabla F(\boldsymbol{x}^+)$ is the gradient at $\boldsymbol{x}^+$, i.e., $\nabla F(\boldsymbol{x}^+)=\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^+-\boldsymbol{y})$. Compare this with GD. Note that HB has two parameters: stepsize $\alpha>0$ and momentum coefficient $\beta\geq 0$. It is known that the iterates of HB satisfy

$$\begin{bmatrix}\boldsymbol{x}^+-\boldsymbol{x}^*\\ \boldsymbol{x}^{++}-\boldsymbol{x}^*\end{bmatrix}=\boldsymbol{N}(\alpha,\beta)\cdot\begin{bmatrix}\boldsymbol{x}-\boldsymbol{x}^*\\ \boldsymbol{x}^+-\boldsymbol{x}^*\end{bmatrix},\text{ where}$$

$$\boldsymbol{N}(\alpha,\beta):=\begin{bmatrix}0 & \boldsymbol{I}\\ -\beta\boldsymbol{I} & (1+\beta)\boldsymbol{I}-\alpha\boldsymbol{A}^\top\boldsymbol{A}\end{bmatrix}.$$

Similarly, we call a stepsize (α, β) optimal if it minimizes $\rho\big(\boldsymbol{N}(\alpha,\beta)\big) =: \rho_{\text{HB}}(\alpha,\beta)$. Let ρ^*_{HB} be the minimum of

$$\min_{\alpha>0,\beta\geq 0} \rho_{\text{HB}}(\alpha,\beta),$$

attained at (α^*, β^*). One can then prove (Polyak, 1964):

$$\rho^*_{\text{HB}} = \frac{\sqrt{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})} - \sqrt{\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})}}{\sqrt{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})} + \sqrt{\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})}},$$
$$\alpha^* = \left(\frac{2}{\sqrt{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})} + \sqrt{\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})}}\right)^2,$$
$$\beta^* = \left(\frac{\sqrt{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})} - \sqrt{\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})}}{\sqrt{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})} + \sqrt{\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})}}\right)^2.$$

2.2. Block Gradient Descent

The phrase "optimal stepsizes" of BGD bears a similar meaning to that of HB. To concretize this, we first rewrite the updates of BGD (let $\boldsymbol{I}$ be the $n \times n$ identity matrix):

Lemma 1. *The iterates of BGD satisfy* $\boldsymbol{x}^+ - \boldsymbol{x}^* = \boldsymbol{M}(\gamma_1,\gamma_2)\cdot(\boldsymbol{x}-\boldsymbol{x}^*)$*, with* $\boldsymbol{M}(\gamma_1,\gamma_2)$ *defined as*

$$\left(\boldsymbol{I} - \begin{bmatrix} 0 & 0 \\ \gamma_2\boldsymbol{A}_2^\top\boldsymbol{A}_1 & \gamma_2\boldsymbol{A}_2^\top\boldsymbol{A}_2 \end{bmatrix}\right)\left(\boldsymbol{I} - \begin{bmatrix} \gamma_1\boldsymbol{A}_1^\top\boldsymbol{A}_1 & \gamma_1\boldsymbol{A}_1^\top\boldsymbol{A}_2 \\ 0 & 0 \end{bmatrix}\right).$$

We call (γ_1,γ_2) optimal if it minimizes $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big) =: \rho_{\text{BGD}}(\gamma_1,\gamma_2)$. Our contribution consists of, under Assumption 1, discovering stepsizes γ_1^*, γ_2^* that satisfy

$$\rho_{\text{BGD}}(\gamma_1^*,\gamma_2^*) = \min_{\gamma_1>0,\gamma_2>0} \rho_{\text{BGD}}(\gamma_1,\gamma_2). \tag{3}$$

With $\rho^*_{\text{BGD}} := \rho_{\text{BGD}}(\gamma_1^*,\gamma_2^*)$, we restate Theorem 1 below:

Theorem 1. *Assumption 1 implies* $\rho^*_{\text{BGD}} \leq (\rho^*_{\text{HB}})^2$.

Vaguely put, a smaller spectral radius implies faster convergence. And the comparison is fair: All stepsizes are chosen to be optimal, minimizing their respective spectral radii; and all methods, namely BGD, GD, and HB, have comparable costs at each iteration. Moreover, $\rho^*_{\text{BGD}} \leq (\rho^*_{\text{HB}})^2$ implies that, with optimal stepsizes, BGD is asymptotically at least twice as fast as HB. Finally, we emphasize it is the presence of Assumption 1 that remains to be interrogated.

3. On Assumption 1

In Section 3.1 We argue Assumption 1 is reasonable, as it relates BGD to (generalized) *alternating projection between two linear subspaces* in such a way that our results under Assumption 1 directly applies and improves existing bounds of Fält & Giselsson (2017) on convergence rates of *generalized alternating projection*. Then, in Section 3.2, we argue Assumption 1 is also essential as it makes analysis possible.

3.1. Justifying Assumption 1

Here we discuss why Assumption 1 is reasonable. Note that it requires the orthogonality of $\boldsymbol{A}_j$ $(j = 1, 2)$, that is $\boldsymbol{A}_j^\top\boldsymbol{A}_j = \boldsymbol{I}_j$, where $\boldsymbol{I}_j$ is the $n_j \times n_j$ identity matrix. First of all, one can always realize this assumption by orthogonalizing $\boldsymbol{A}_1, \boldsymbol{A}_2$ respectively (even though it entails some computational costs). Secondly, and more importantly, this block-wise orthogonality assumption bridges BGD and the classic method of alternating projection between two subspaces. To see this, let us consider the following lemma:

Lemma 2. *The iterates* $\boldsymbol{x}^+, \boldsymbol{x}$ *of BGD satisfy* $\boldsymbol{A}(\boldsymbol{x}^+ - \boldsymbol{x}^*) = \boldsymbol{T}(\gamma_1,\gamma_2)\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}^*)$*, with* $\boldsymbol{T}(\gamma_1,\gamma_2)$ *defined as*

$$\boldsymbol{T}(\gamma_1,\gamma_2) := \left(\boldsymbol{I} - \gamma_2\boldsymbol{A}_2\boldsymbol{A}_2^\top\right)\left(\boldsymbol{I} - \gamma_1\boldsymbol{A}_1\boldsymbol{A}_1^\top\right).$$

Lemma 2 implies BGD can be viewed as an algorithm that operates on the iterates $\boldsymbol{z}^+ := \boldsymbol{A}(\boldsymbol{x}^+ - \boldsymbol{x}^*)$ and $\boldsymbol{z} := \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*)$. Since $\boldsymbol{A}$ is assumed to be full column rank, updating $\boldsymbol{z}^+$ is equivalent to updating $\boldsymbol{x}^+$. If Assumption 1 holds, then $\boldsymbol{I} - \boldsymbol{A}_j\boldsymbol{A}_j^\top$ $(j = 1, 2)$ is an orthogonal projection, the update $\boldsymbol{z}^+ = \boldsymbol{T}(1,1)\boldsymbol{z}$ is precisely the vanilla alternating projection method between two subspaces (von Neumann, 1951), and the update $\boldsymbol{z}^+ = \boldsymbol{T}(\gamma_1,\gamma_2)\boldsymbol{z}$ can be viewed as *generalized* alternating projection (Fält & Giselsson, 2017).

With the above reasoning, we can intuitively conclude that studying the convergence of BGD under Assumption 1 would also yield convergence guarantees for generalized alternating projection. A more detailed treatment from the perspective of alternating projection can be found in Section 5.5, where we rectify the slightly misleading claim of Fält & Giselsson (2017) that their bound was optimal.

Finally, Assumption 1 requires $\boldsymbol{A}_2^\top\boldsymbol{A}_1 \neq 0$; this is to sidestep the trivial case where BGD, GD, and HB all converge to $\boldsymbol{x}^*$ in just 1 iteration with appropriate stepsizes under the block-wise orthogonality assumption.

3.2. The Technical Role of Assumption 1

Note that $\boldsymbol{M}(\gamma_1,\gamma_2)$ has a sophisticated expression, and so Assumption 1 plays the role of simplifying, at least making analyzing (3) possible. In particular, Assumption 1 immediately simplifies the expression of $\boldsymbol{M}(\gamma_1,\gamma_2)$:

Lemma 3. *Define* $\boldsymbol{C} := \boldsymbol{A}_2^\top\boldsymbol{A}_1$*. Recall the definition of* $\boldsymbol{M}(\gamma_1,\gamma_2)$ *in Lemma 1. Assumption 1 implies*

$$\boldsymbol{M}(\gamma_1,\gamma_2) = \begin{bmatrix} (1-\gamma_1)\boldsymbol{I}_1 & -\gamma_1\boldsymbol{C}^\top \\ -\gamma_2(1-\gamma_1)\boldsymbol{C} & (1-\gamma_2)\boldsymbol{I}_2 + \gamma_1\gamma_2\boldsymbol{C}\boldsymbol{C}^\top \end{bmatrix}.$$

We note that Assumption 1 does not clean all obstacles, as $\boldsymbol{M}(\gamma_1,\gamma_2)$ in Lemma 3 is still complicated, e.g., it depends on γ_1, γ_2 quadratically; to compare, the corresponding matrix $\boldsymbol{N}(\alpha,\beta)$ of HB has a linear dependency on α, β.

After a simplification through Assumption 1, $\boldsymbol{M}(\gamma_1, \gamma_2)$ still maintains an interesting structure that will ultimately facilitate our understanding of BGD. For example, $\boldsymbol{M}(\gamma_1, \gamma_2)$ depends on $\boldsymbol{C}$, and $\boldsymbol{C}$ is precisely the matrix that encodes the information about the relationship between the two blocks of variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. It is by leveraging the spectrum of $\boldsymbol{C}$, and therefore of $\boldsymbol{M}(\gamma_1, \gamma_2)$, that we will be able to show BGD enjoys faster convergence than its competitors GD and HB. To get prepared for the competition, we use Assumption 1 and express the minimum spectral radii ρ^*_{HB} and ρ^*_{GD} in terms of the spectrum of $\boldsymbol{C}$:

Lemma 4. *Recall $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2] \in \mathbb{R}^{m\times(n_1+n_2)}$, $\boldsymbol{A}$ is full rank, and $\boldsymbol{C} = \boldsymbol{A}_2^\top \boldsymbol{A}_1 \in \mathbb{R}^{n_2\times n_1}$. Suppose $\operatorname{rank}(\boldsymbol{C}) = r$ and Assumption 1 holds. Then $\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) \neq 1$ and the maximum and minimum eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ are $1+\sqrt{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}$ and $1-\sqrt{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}$, respectively. Hence, the minimum spectral radii of GD and HB are given as*

$$\rho^*_{\text{GD}} = \sqrt{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)},\ \rho^*_{\text{HB}} = \frac{1-\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}{\sqrt{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}. \quad (4)$$

Example 1. With Lemma 4 and Assumption 1, we have

$$\rho\big(\boldsymbol{M}(1,1)\big) = \lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) < \sqrt{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)} = \rho^*_{\text{GD}},$$

meaning BEM converges asymptotically faster than GD.

Is it possible that $\gamma_1 = \gamma_2 = 1$ actually minimizes $\rho\big(\boldsymbol{M}(\gamma_1, \gamma_2)\big)$? After all, BEM already ensures the largest possible decrease of the objective for the present block, while all other stepsizes guarantee less! If that were true, then $\rho\big(\boldsymbol{M}(1,1)\big)$ would not necessarily be smaller than ρ^*_{HB}, let alone our promise $\rho^*_{\text{BGD}} \leq (\rho^*_{\text{HB}})^2$. The sole hope is that other stepsizes, though sub-optimal for the moment, might be more beneficial in the long run—and if so, we simply need to work harder to find them.

4. Optimal Stepsizes and Spectral Radius

The inequality $\rho^*_{\text{BGD}} \leq (\rho^*_{\text{HB}})^2$ and hence Theorem 1 are proved under Assumption 1 by deriving a closed-form expression of the minimum value ρ^*_{BGD} of (3) and the corresponding stepsizes. The derivation is summarized below:

Theorem 2. *Recall $\rho^*_{\text{BGD}} = \rho_{\text{BGD}}(\gamma_1^*, \gamma_2^*)$ in (3) and $\boldsymbol{C} := \boldsymbol{A}_2^\top \boldsymbol{A}_1$. Assumption 1 implies the following.*

- *If $\boldsymbol{C}$ has full rank (i.e., $\operatorname{rank}(\boldsymbol{C}) = \min\{n_1, n_2\}$), then*

$$\rho^*_{\text{BGD}} = \frac{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)} - \sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)} + \sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}. \quad (5)$$

- *If $\boldsymbol{C}$ is rank-deficient, then*

$$\rho^*_{\text{BGD}} = \frac{1-\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}{1+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}. \quad (6)$$

and the corresponding stepsizes (γ_1^*, γ_2^*) are given as

$$\gamma_1^* = \gamma_2^* = \frac{2}{1+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}. \quad (7)$$

Remark 1. If $r = 1$, then we have $\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) = \lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)$ and Theorem 2 suggests the minimum spectral radius is 0. When $\boldsymbol{C}$ has full rank, the minimizers of $\rho\big(\boldsymbol{M}(\gamma_1, \gamma_2)\big)$ have complicated forms, so we do not show them here; they can be found in the proof of Appendix C.1.

Remark 2. The full rank case exhibits a smaller minimum spectral radius than the rank-deficient case, and the minimum spectral radius in the rank-deficient case is precisely $(\rho^*_{\text{HB}})^2$, where ρ^*_{HB} is defined in (4). Explicitly, we have

$$\begin{aligned}(5) &= \frac{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)} - \sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)} + \sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}} \\ &\leq \frac{1-\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}{1+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}} = (6) = (\rho^*_{\text{HB}})^2.\end{aligned}$$

This consolidates Theorem 1, and implies, under Assumption 1, BGD is eventually twice as fast as HB.

Remark 3. Interestingly, the optimal stepsizes (γ_1^*, γ_2^*) derived in (7) are the same as the optimal stepsize α^* of HB. Indeed, with Assumption 1 and Lemma 4, we have

$$\begin{aligned}\gamma_1^* = \gamma_2^* &= \frac{2}{1+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}} \\ &= \left(\frac{2}{\sqrt{\lambda_1(\boldsymbol{A}^\top\boldsymbol{A})} + \sqrt{\lambda_n(\boldsymbol{A}^\top\boldsymbol{A})}}\right)^2 = \alpha^*.\end{aligned}$$

This offers some intuition as to why BGD is *twice* as fast under Assumption 1: BGD and HB take the same stepsize, but BGD takes *two* descent steps—with partial gradients and without momentum—in a single iteration, entailing a cost barely comparable to one descent step of HB.

5. Related Work

5.1. BGD Versus GD and HB

Theorem 1 shows that BGD can be faster than the heavy ball method (HB), and the latter has thus far been one of the theoretically fastest variants of accelerated GD for least-squares. Interestingly, though, BGD has no momentum at all! In hindsight, our justification is that BGD is a *two-step* method in the sense of Polyak (1987): It has two stepsizes γ_1, γ_2, comparable to the stepsize and momentum coefficient in HB, so, similarly to HB, choosing the stepsizes γ_1, γ_2 in an "optimal" way results in acceleration. While several methods that accelerate without momentum exist (e.g., Young's method (Young, 1953), GD with cyclic stepsizes, see Section 5.2), our approach is unique in the sense that it achieves acceleration by operating on *partial* gradients without momentum in a cyclic, block-wise manner.

5.2. BGD Versus GD with Cyclic Stepsizes

There has been some recent interest in using GD with stepsizes selected in a cyclic fashion from a sequence of M positive numbers (Smith, 2017; Oymak, 2021; Goujaud et al., 2022; Grimmer, 2023); see also *fractal* stepsizes of Agarwal et al. (2021) and silver stepsizes of Altschuler & Parrilo (2023a;b). While Grimmer (2023) and Altschuler & Parrilo (2023a;b) considered GD in the more general setting of (strongly-)convex smooth optimization, more relevant to ours is the work of Oymak (2021); Goujaud et al. (2022). Oymak (2021) analyzed GD with cyclic stepsizes in the least-squares context and Goujaud et al. (2022) analyzed HB with cyclic stepsizes. Both Oymak (2021) and Goujaud et al. (2022) assume the spectrum of $\boldsymbol{A}^\top \boldsymbol{A}$ is *clustered* into two or more disjoint intervals; this is different from Assumption 1, so the results of Oymak (2021); Goujaud et al. (2022) are not directly comparable to ours. It should be noted that the use of cyclic stepsizes is inherent in BGD. In fact, BGD not only cycles between stepsizes, but also between blocks. One more difference is this: BGD cycles *more frequently*, meaning that it leverages two stepsizes in a single iteration with a cost comparable to one GD step, while GD with cyclic stepsizes needs M iterations to make full use of M stepsizes.

5.3. BGD Versus Lower Bounds

As shown in Assumption 2.1.4 & Theorem 2.1.13 of Nesterov (2018b), for constants $L > \mu > 0$, there is a μ-strongly convex and L-smooth function $F(\boldsymbol{x})$, globally minimized at $\boldsymbol{x}^*$, so that the iterates $\boldsymbol{x}^t$ of any first-order method created from linear combinations of any initialization $\boldsymbol{x}^0$ and previous gradient evaluations $\nabla F(\boldsymbol{x}^0), \dots, \nabla F(\boldsymbol{x}^{t-1})$ satisfy the lower bound

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t \cdot \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2, \quad \kappa := L/\mu. \tag{8}$$

See also Theorem 3.1 of Sun & Ye (2021), which proves a similar lower bound for the multi-block version of BGD with block sizes 1 and canonical stepsizes $1/L$ applied to least-squares. While HB asymptotically attains lower bound (8), our Theorem 1 suggests that, under Assumption 1, the iterates $\boldsymbol{x}^t$ of BGD with optimal stepsizes would satisfy

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \cdot \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2$$

asymptotically (as $t \to \infty$). Is this a contradiction or does that mean BGD breaches lower bound (8)? The answer is no and the reason is two-fold. First, BGD is not a first-order method in the precise sense of Assumption 2.1.4 of Nesterov (2018b), e.g., its new iterates do not arise as linear combinations of previous ones or their gradients. Second, the objective function we optimize is least-squares with Assumption 1, which does not necessarily belong to the class of worst-case functions in Theorem 2.1.13 of Nesterov (2018b) or in Theorem 3.1 of Sun & Ye (2021). One more subtlety is that the lower bound of Sun & Ye (2021) is derived for *canonical*, rather than optimal, stepsizes.

5.4. BGD Versus Block Exact Minimization

Under Assumption 1, one verifies BGD with stepsizes $\gamma_1 = 1, \gamma_2 = 1$ is equivalent[2] to *block exact minimization* (BEM),

$$\begin{aligned} \boldsymbol{x}_1^+ &\in \operatorname*{argmin}_{\boldsymbol{x}_1 \in \mathbb{R}^{n_1}} F([\boldsymbol{x}_1; \boldsymbol{x}_2]), \\ \boldsymbol{x}_2^+ &\in \operatorname*{argmin}_{\boldsymbol{x}_2 \in \mathbb{R}^{n_2}} F([\boldsymbol{x}_1^+; \boldsymbol{x}_2]), \end{aligned} \tag{BEM}$$

where F is the least-squares objective in (2), so BGD under Assumption 1 generalizes BEM by allowing for different stepsizes. While BEM ensures the largest possible decrease of F when updating either $\boldsymbol{x}_1^+$ or $\boldsymbol{x}_2^+$, we have seen the optimal stepsizes of BGD are not as simple as $\gamma_1 = 1, \gamma_2 = 1$; BGD can do better than BEM under Assumption 1.

5.5. BGD and Generalized Alternating Projection

As shown in Lemma 2, BGD is highly related to alternating projection between two subspaces, to which our results, say Theorem 2, apply. Hence, in this subsection, we review existing theoretical results on alternating projection and its variants, and then later in Section 6 we show how our Theorem 2 applies and improves prior work.

Given two linear subspaces $\mathcal{H}_1, \mathcal{H}_2$ of $\mathbb{R}^m$ of dimension $m - n_1, m - n_2$ respectively, we are interested in projecting some point $\boldsymbol{z}^{(0)} \in \mathbb{R}^m$ onto the intersection $\mathcal{H}_1 \cap \mathcal{H}_2$. To avoid trivial cases, we assume $\mathcal{H}_1$ does not contain $\mathcal{H}_2$ (and vice versa), and none of the two subspaces is $\{0\}$.

We consider iterative algorithms of the form

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{T}\boldsymbol{z}^{(t)}, \tag{9}$$

where $\boldsymbol{T} \in \mathbb{R}^{m \times m}$ is some matrix that transforms the current iterate $\boldsymbol{z}^{(t)}$ into the next, $\boldsymbol{z}^{(t+1)}$. The choice of $\boldsymbol{T}$ typically depends on the two subspaces $\mathcal{H}_1, \mathcal{H}_2$. We next review several algorithms that differ in how $\boldsymbol{T}$ is chosen and that converge at different rates (see Table 1 for a summary).

5.5.1. Project Versus Reflect

Let us first review two classic methods to project a point onto $\mathcal{H}_1 \cap \mathcal{H}_2$: alternating projection and Douglas-Rachford.

Alternating Projection. Let $\boldsymbol{P}_{\mathcal{H}_1}, \boldsymbol{P}_{\mathcal{H}_2}$ be matrices representing orthogonal projections onto $\mathcal{H}_1, \mathcal{H}_2$ respectively.

[2] Under Assumption 1, the Lipschitz smoothness constant L_1 (or L_2) of (2) in variable $\boldsymbol{x}_1$ (or $\boldsymbol{x}_2$) is 1, so stepsizes $(\gamma_1, \gamma_2) = (1, 1)$ also correspond to the commonly used choice $(1/L_1, 1/L_2)$.

The *alternating projection* algorithm sets $\boldsymbol{T}$ to be

$$\boldsymbol{T} := \boldsymbol{P}_{\mathcal{H}_2}\boldsymbol{P}_{\mathcal{H}_1}. \tag{AP}$$

It follows immediately that for every t we have

$$\|\boldsymbol{z}^{(t)} - \boldsymbol{P}_{\mathcal{H}_1\cap\mathcal{H}_2}\boldsymbol{z}^{(0)}\|_2 \le \|\boldsymbol{T}^k - \boldsymbol{P}_{\mathcal{H}_1\cap\mathcal{H}_2}\|_2 \cdot \|\boldsymbol{z}^{(0)}\|_2.$$

As Kayalar & Weinert (1988) reviewed, Von Neumann established that AP converges to $\boldsymbol{P}_{\mathcal{H}_1\cap\mathcal{H}_2}\boldsymbol{z}^{(0)}$ in the early 1930s and his result was not in print until almost two decades after; see, e.g., Theorem 13.7 of von Neumann (1951). Around the same time, Aronszajn (1950) showed $\boldsymbol{T}$ of AP satisfies

$$\|\boldsymbol{T}^k - \boldsymbol{P}_{\mathcal{H}_1\cap\mathcal{H}_2}\|_2 \le \cos^{2k-1}\theta_1, \tag{10}$$

where $\theta_1 \in [0, \pi/2]$ is defined to be the minimum nonzero *principal angle* between $\mathcal{H}_1$ and $\mathcal{H}_2$, also widely known as the Friedrichs angle (Friedrichs, 1937).[3] While Deutsch (1984); Franchetti & Light (1986) announced bounds that are looser than (10), Theorem 2 of Kayalar & Weinert (1988) proved that (10) always holds with equality, which implies this linear rate $\cos^2\theta_1$ is actually optimal for AP.

Douglas-Rachford. Another classic iterative scheme proposed (implicitly) by Douglas & Rachford (1956) is this:

$$\begin{aligned}\boldsymbol{T} &:= \boldsymbol{P}_{\mathcal{H}_2}(2\boldsymbol{P}_{\mathcal{H}_1} - \boldsymbol{I}) + \boldsymbol{I} - \boldsymbol{P}_{\mathcal{H}_1} \\ &= \frac{1}{2}\boldsymbol{I} + \frac{1}{2}(2\boldsymbol{P}_{\mathcal{H}_2} - \boldsymbol{I})(2\boldsymbol{P}_{\mathcal{H}_1} - \boldsymbol{I}).\end{aligned} \tag{DR}$$

In words, DR consists of applying in cascade two reflections, $2\boldsymbol{P}_{\mathcal{H}_1} - \boldsymbol{I}$ and then $2\boldsymbol{P}_{\mathcal{H}_2} - \boldsymbol{I}$, to the current iterate $\boldsymbol{z}^{(t)}$ and then taking an average. Hesse et al. (2014) showed in their Theorem IV.6 that DR converges linearly without providing any explicit rate, while Bauschke et al. (2014) later made the rate precise: Their Theorem 4.3 proves $\boldsymbol{T}$ of DR satisfies

$$\|\boldsymbol{P}_{\mathcal{H}_1}\boldsymbol{T}^k - \boldsymbol{P}_{\mathcal{H}_1\cap\mathcal{H}_2}\|_2 = \|\boldsymbol{P}_{\mathcal{H}_2}\boldsymbol{T}^k - \boldsymbol{P}_{\mathcal{H}_1\cap\mathcal{H}_2}\|_2 = \cos^k\theta_1.$$

We can then say that DR converges linearly with rate $\cos\theta_1$ and this is optimal for DR. Note though that since $\cos^2\theta_1 \le \cos\theta_1$, DR is provably slower than AP.

5.5.2. Four Generalized Methods

With two positive numbers γ_1, γ_2 and the $m \times m$ identity matrix $\boldsymbol{I}$, define the relaxed projections $\boldsymbol{P}^{\gamma_1}_{\mathcal{H}_1}$ and $\boldsymbol{P}^{\gamma_2}_{\mathcal{H}_2}$ as

$$\begin{aligned}\boldsymbol{P}^{\gamma_1}_{\mathcal{H}_1} &:= (1-\gamma_1)\boldsymbol{I} + \gamma_1\boldsymbol{P}_{\mathcal{H}_1}, \\ \boldsymbol{P}^{\gamma_2}_{\mathcal{H}_2} &:= (1-\gamma_2)\boldsymbol{I} + \gamma_1\boldsymbol{P}_{\mathcal{H}_2}.\end{aligned} \tag{11}$$

[3] Alternating projection is also related to a popular line of research called *continual learning* (Peng et al., 2023; Elenter et al., 2023; Cai & Diakonikolas, 2024). There, convergence results similar to (10) can be found; see, e.g., Theorem 8 of Evron et al. (2022). Our stepsizes and theory can be applied there to improve the rate $\cos^2(\theta_1)$ in Theorem 8 of Evron et al. (2022) to $\frac{\sin(\theta_r)-\sin(\theta_1)}{\sin(\theta_r)+\sin(\theta_1)}$. Here, θ_r and θ_1 denote the largest and nonzero smallest principal angles between $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. See also Table 1.

Note that $\boldsymbol{P}^1_{\mathcal{H}_1}$ represents the projection $\boldsymbol{P}_{\mathcal{H}_1}$ and $\boldsymbol{P}^2_{\mathcal{H}_1}$ the reflection, so $\boldsymbol{P}^{\gamma_1}_{\mathcal{H}_1}$ generalizes the operators in AP and DR. For clarity, we will always write $\boldsymbol{P}^{\gamma_1}_{\mathcal{H}_1}$ to mean definitions in (11) and it never means the matrix $\boldsymbol{P}_{\mathcal{H}_1}$ raised to the power of γ_1; if ever needed, we will write $(\boldsymbol{P}_{\mathcal{H}_1})^{\gamma_1}$ for the latter.

With the relaxed projections $\boldsymbol{P}^{\gamma_1}_{\mathcal{H}_1}$ and $\boldsymbol{P}^{\gamma_2}_{\mathcal{H}_2}$, we can now proceed to review methods that generalize AP and DR.

Relaxed Alternating Projection. For some $\gamma \in (0, 1]$, the method we review now has the iterates $\boldsymbol{z}^{(t+1)} = \boldsymbol{T}(\gamma)\boldsymbol{z}^{(t)}$, where $\boldsymbol{T}(\gamma)$ is a matrix function of γ, defined as

$$\boldsymbol{T}(\gamma) := (1-\gamma)\boldsymbol{I} + \gamma\boldsymbol{P}_{\mathcal{H}_2}\boldsymbol{P}_{\mathcal{H}_1}. \tag{RAP}$$

Clearly, $\boldsymbol{T}(1)$ corresponds to AP. How to choose γ to obtain a faster rate than AP? The optimal choice of γ would be 1 if $\boldsymbol{P}_{\mathcal{H}_2}\boldsymbol{P}_{\mathcal{H}_1}$ were symmetric, while it is precisely the asymmetry of $\boldsymbol{P}_{\mathcal{H}_2}\boldsymbol{P}_{\mathcal{H}_1}$ that brings difficulty. Bauschke et al. (2016) showed in their Theorem 3.6 that the choice $\gamma = \frac{2}{1+\sin^2(\theta_1)}$ is optimal, with which RAP converges asymptotically at a linear rate $\frac{1-\sin^2(\theta_1)}{1+\sin^2(\theta_1)}$ faster than AP.

Partial Relaxed Alternating Projection. The method we present now relaxes AP *partially* by setting

$$\boldsymbol{T}(\gamma_1) := \boldsymbol{P}_{\mathcal{H}_2}\boldsymbol{P}^{\gamma_1}_{\mathcal{H}_1}. \tag{PRAP}$$

Theorem 3.7 of Bauschke et al. (2016) proves the optimal choice of γ_1 for PRAP is $\frac{2}{\sin^2(\theta_r)+\sin^2(\theta_1)}$, giving the asymptotic linear rate $\frac{\sin^2(\theta_r)-\sin^2(\theta_1)}{\sin^2(\theta_r)+\sin^2(\theta_1)}$. Moreover, one verifies this rate is faster than or the same as those of AP, DR, and RAP.

Generalized Douglas-Rachford. In their Section 1.4, Demanet & Zhang (2016) considered the following operator $T(\gamma)$ with iterate update $z^{k+1} = \boldsymbol{T}(\gamma)z^k$:

$$\boldsymbol{T}(\gamma) = (1-\gamma)\boldsymbol{I} + \gamma\left(\frac{1}{2}\boldsymbol{I} + \frac{1}{2}\boldsymbol{P}^2_{\mathcal{H}_2}\boldsymbol{P}^2_{\mathcal{H}_1}\right). \tag{GDR}$$

GDR can be viewed as taking a convex combination of $\boldsymbol{I}$ and the operator of DR, similarly to RAP and AP. However, choosing $\gamma \in [0, 1]$ for GDR does not improve the rate: Theorem 3.10 of Bauschke et al. (2016) shows $\gamma = 1$ is optimal for GDR, giving the same linear rate $\cos(\theta_1)$ as DR.

Generalized Alternating Projection. The generalized alternating projection algorithm is with the operator

$$\boldsymbol{T}(\gamma, \gamma_1, \gamma_2) = (1-\gamma)\boldsymbol{I} + \gamma\boldsymbol{P}^{\gamma_2}_{\mathcal{H}_2}\boldsymbol{P}^{\gamma_1}_{\mathcal{H}_1}. \tag{GAP}$$

where γ is assumed to lie in $(0, 1]$. If γ and γ_2 were taken to be 1, then GAP recovers PRAP. Fält & Giselsson (2017) proposed the stepsizes $\gamma = 1, \gamma_1 = \gamma_2 = \frac{2}{1+\sin(\theta_1)}$, with which GAP converges asymptotically at linear rate $\frac{1-\sin(\theta_1)}{1+\sin(\theta_1)}$. Moreover, Fält & Giselsson (2017) shows this choice of stepsizes is optimal as long as the largest principal angle θ_r between the two subspace $\mathcal{H}_1, \mathcal{H}_2$ is equal to $\pi/2$. Indeed, if θ_r were equal to $\pi/2$, then the rate $\frac{1-\sin(\theta_1)}{1+\sin(\theta_1)}$ of Fält & Giselsson (2017) is the smallest among prior methods.

Table 1: Comparison of GAP++ (Section 6) with existing rates. A smaller linear rate means better, and θ_r and θ_1 denote the largest and nonzero smallest principal angles between $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. We find novel stepsizes (GAP++) that guarantee faster convergence. (The stepsizes γ_1^*, γ_2^* of GAP++ are shown in Remark 12 of the appendix.)

Method	Linear Rate Factor	Stepsizes	Asymptotic?	Reference
AP	$\cos^2(\theta_1)$	N.A.	No	Theorem 2 (Kayalar & Weinert, 1988)
DR	$\cos(\theta_1)$	N.A.	No	Theorem 4.3 (Bauschke et al., 2014)
RAP	$\frac{1-\sin^2(\theta_1)}{1+\sin^2(\theta_1)}$	$\gamma = \frac{2}{1+\sin^2(\theta_1)}$	Yes	Theorem 3.6 (Bauschke et al., 2016)
PRAP	$\frac{\sin^2(\theta_r)-\sin^2(\theta_1)}{\sin^2(\theta_r)+\sin^2(\theta_1)}$	$\gamma_1 = \frac{2}{\sin^2(\theta_r)+\sin^2(\theta_1)}$	Yes	Theorem 3.7 (Bauschke et al., 2016)
GDR	$\cos(\theta_1)$	$\gamma = 1$	No	Theorem 3.10 (Bauschke et al., 2016)
GAP	$\frac{1-\sin(\theta_1)}{1+\sin(\theta_1)}$	$\gamma = 1, \gamma_1 = \gamma_2 = \frac{2}{1+\sin(\theta_1)}$	Yes	Theorem 3, Remark 3 (Fält & Giselsson, 2017)
GAP++	$\frac{\sin(\theta_r)-\sin(\theta_1)}{\sin(\theta_r)+\sin(\theta_1)}$	$\gamma = 1, \gamma_1 = \gamma_1^*, \gamma_2 = \gamma_2^*$	Yes	This Paper, Corollary 1

5.5.3. The Gap

In light of the review in Section 5.5.2, a major gap to be bridged is as follows. The assumption $\theta_r = \pi/2$ of Fält & Giselsson (2017) on the largest principal angle θ_r is violated with probability 1 by any two subspaces of fixed dimensions randomly sampled from their respective Grassmannian manifolds; this means their stepsizes are sub-optimal with probability 1. This sub-optimality manifests itself in comparison to PRAP: The stepsizes of Fält & Giselsson (2017) do not necessarily yield a smaller rate than PRAP of Bauschke et al. (2016). More specifically, the rate $\frac{\sin^2(\theta_r)-\sin^2(\theta_1)}{\sin^2(\theta_r)+\sin^2(\theta_1)}$ of PRAP could be larger or smaller than the rate $\frac{1-\sin(\theta_1)}{1+\sin(\theta_1)}$ of Fält & Giselsson (2017), depending on problem configurations (e.g., the precise values of θ_r).

6. GAP++: Bridge the GAP

A major contribution of this section is closing the gap discussed in Section 5.5.3. This is achieved by invoking our theoretical results developed so far. Specifically, with the stepsizes that we propose, the (GAP) operator converges asymptotically at rate $\frac{\sin(\theta_r)-\sin(\theta_1)}{\sin(\theta_r)+\sin(\theta_1)}$. This rate is alway better than if not the same as the rates of Bauschke et al. (2016) and Fält & Giselsson (2017), for all possible choices of θ_1 and θ_r. On the conceptual level, our result can be viewed as a generalization of Fält & Giselsson (2017) which dispenses with their assumption $\theta_r = \pi/2$. Therefore we call the proposed stepsizes GAP++.

In Section 6.1 we build preliminary notations. In Section 6.2 we provide stepsizes that achieve the fastest rate in Table 1.

6.1. Preliminaries

Matrix Representations. For $j = 1, 2$, let $A_j \in \mathbb{R}^{m\times n_j}$ be an orthonormal basis matrix of the orthogonal complement $\mathcal{H}_j^\perp$ of $\mathcal{H}_j$. Then we can write $\boldsymbol{P}_{\mathcal{H}_j} := \boldsymbol{I} - A_j A_j^\top$ and the relaxed projections $\boldsymbol{P}_{\mathcal{H}_1}^{\gamma_2}, \boldsymbol{P}_{\mathcal{H}_1}^{\gamma_2}$ in (11) as

$$\boldsymbol{P}_{\mathcal{H}_j}^{\gamma_j} = \boldsymbol{I} - \gamma_j A_j A_j^\top, \quad j = 1, 2. \tag{12}$$

Note that, here, we write the symbols A_1, A_2 without boldface to distinguish them from the notations $\boldsymbol{A}_1, \boldsymbol{A}_2$ of the main paper. However, the choices of the same letter A, $\boldsymbol{A}$ should remind the reader that they will eventually be connected somehow.

The GAP operator can be expressed in terms of A_1, A_2:

$$\boldsymbol{T}(\gamma, \gamma_1, \gamma_2) = (1-\gamma)\boldsymbol{I} + \gamma(\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top).$$

We will set $\gamma = 1$ and analyze $\boldsymbol{T}(1, \gamma_1, \gamma_2) = (\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)$, as doing so is sufficient to derive a faster rate than that of Fält & Giselsson (2017).

Principal Angles. We recall the following definition of *nonzero* principal angles.

Definition 1. For subspaces $\mathcal{H}_1, \mathcal{H}_2$ of $\mathbb{R}^m$, define $r := \min\{\dim(\mathcal{H}_1), \dim(\mathcal{H}_2)\} - \dim(\mathcal{H}_1 \cap \mathcal{H}_2)$ and $\mathcal{S}_1 := \mathcal{H}_1 \cap (\mathcal{H}_1 \cap \mathcal{H}_2)^\perp$ and $\mathcal{Z}_1 := \mathcal{H}_2 \cap (\mathcal{H}_1 \cap \mathcal{H}_2)^\perp$. The *nonzero principal angles* $\theta_1, \ldots, \theta_r$ between $\mathcal{H}_1$ and $\mathcal{H}_2$ are then defined recursively for $i = 1, \ldots, r$ such that

$$(\boldsymbol{s}_i, \boldsymbol{z}_i) \in \underset{\substack{\boldsymbol{s}\in\mathcal{S}_i, \boldsymbol{z}\in\mathcal{Z}_i \\ \|\boldsymbol{s}\|_2 = \|\boldsymbol{z}\|_2 = 1}}{\operatorname{argmax}} \boldsymbol{s}^\top \boldsymbol{z}, \quad \theta_i = \arccos(\boldsymbol{s}_i^\top \boldsymbol{z}_i), \tag{13}$$

where $\mathcal{S}_i := \mathcal{S}_{i-1} \cap \operatorname{span}(\boldsymbol{s}_i)^\perp$, $\mathcal{Z}_i := \mathcal{Z}_{i-1} \cap \operatorname{span}(\boldsymbol{z}_i)^\perp$.

It is not hard to show the principal angles $\theta_1, \ldots, \theta_r$ between $\mathcal{H}_1$ and $\mathcal{H}_2$ are indeed nonzero and satisfy $\theta_1 \le \cdots \le \theta_r$. Moreover, these principal angles between $\mathcal{H}_1, \mathcal{H}_2$ correspond exactly to nonzero principal angles between the orthogonal complements $\mathcal{H}_1^\perp$ and $\mathcal{H}_2^\perp$; see Property 2.1 of Zhu & Knyazev (2013) or Theorem 2.7 of Knyazev & Argentati (2007). More formally, we have the following statements:

Lemma 5. *Recall that $A_j \in \mathbb{R}^{m\times n_j}$ is an orthonormal basis matrix of the orthogonal complement $\mathcal{H}_j^\perp$ of $\mathcal{H}_j$ ($j = 1, 2$), and that $r := \min\{\dim(\mathcal{H}_1), \dim(\mathcal{H}_2)\} - \dim(\mathcal{H}_1 \cap$*

$\mathcal{H}_2$). *Define* $\boldsymbol{C} := A_2^\top A_1$. *Then we have* $\mathrm{rank}(\boldsymbol{C}) = r$. *Denote by* $\sigma_1 \geq \cdots \geq \sigma_r$ *the* r *nonzero singular values of* $\boldsymbol{C}$. *Then* $\sigma_i = \cos(\theta_i)$ *for every* $i = 1, \ldots, r$.

6.2. Asymptotic Convergence Rates

Here we give stepsize choices γ_1, γ_2 and the associated asymptotic convergence rates for the iterates $\boldsymbol{z}^+ = \boldsymbol{T}(1, \gamma_1, \gamma_2)\boldsymbol{z}$, where we recall $\boldsymbol{T}(1, \gamma_1, \gamma_2)$ is defined as $\boldsymbol{T}(1, \gamma_1, \gamma_2) = (\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)$.

Write $A := [A_1 \ A_2] \in \mathbb{R}^{m \times (n_1+n_2)}$ and assume A has full column rank (necessarily, we assume $m \geq n_1 + n_2 =: n$). Let $\boldsymbol{Q}$ be an orthonormal basis matrix for $\mathcal{H}_1 \cap \mathcal{H}_2$. Then $\boldsymbol{Q}^\top \boldsymbol{Q}$ is the identity matrix, and we have $\boldsymbol{Q}^\top A = 0$. Furthermore, we can write $\boldsymbol{z}$ as $\boldsymbol{z} = A\boldsymbol{c}_1 + \boldsymbol{Q}\boldsymbol{c}_2$ and then

$$\begin{aligned}
&\boldsymbol{T}(1, \gamma_1, \gamma_2)\boldsymbol{z} = \boldsymbol{T}(1, \gamma_1, \gamma_2)(A\boldsymbol{c}_1 + \boldsymbol{Q}\boldsymbol{c}_2) \\
&= (\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)(A\boldsymbol{c}_1 + \boldsymbol{Q}\boldsymbol{c}_2) \\
&= (\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)A\boldsymbol{c}_1 + \boldsymbol{Q}\boldsymbol{c}_2.
\end{aligned}$$

Since the columns of $\boldsymbol{Q}$ span $\mathcal{H}_1 \cap \mathcal{H}_2$ and the goal is to reach a point in $\mathcal{H}_1 \cap \mathcal{H}_2$, we could assume $\boldsymbol{c}_2 = 0$ without loss of generality, and analyze the update equation

$$\boldsymbol{d} := \boldsymbol{T}(1, \gamma_1, \gamma_2)A\boldsymbol{c}_1 = (\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)A\boldsymbol{c}_1.$$

Moreover, since A has full column rank and $\boldsymbol{Q}^\top \boldsymbol{d} = \boldsymbol{Q}^\top(\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)A\boldsymbol{c}_1 = 0$, the new iterate $\boldsymbol{d}$ always lies in the range space of A, meaning that we can uniquely write $\boldsymbol{d}$ as $\boldsymbol{d} = A\boldsymbol{c}_1^+$ for some $\boldsymbol{c}_1^+$. Then the update equation that we need to analyze become

$$\begin{aligned}
&A\boldsymbol{c}_1^+ = (\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)A\boldsymbol{c}_1 \qquad (14) \\
\Leftrightarrow\ &\boldsymbol{c}_1^+ = (A^\top A)^{-1} A^\top (\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)A\boldsymbol{c}_1.
\end{aligned}$$

We need the following lemma to proceed.

Lemma 6. *If* $A \in \mathbb{R}^{m \times (n_1+n_2)}$ *has full column rank, then the matrix* $(A^\top A)^{-1} A^\top (\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)A$ *is equal to* $M(\gamma_1, \gamma_2)$, *where* $M(\gamma_1, \gamma_2)$ *is defined to be*

$$\left(\boldsymbol{I} - \begin{bmatrix} 0 & 0 \\ \gamma_2 A_2^\top A_1 & \gamma_2 \boldsymbol{I}_2 \end{bmatrix}\right)\left(\boldsymbol{I} - \begin{bmatrix} \gamma_1 \boldsymbol{I}_1 & \gamma_1 A_1^\top A_2 \\ 0 & 0 \end{bmatrix}\right).$$

With (14) and Lemma 6, we conclude that we need to analyze the convergence rate of the iterate update

$$\boldsymbol{c}_1^+ = M(\gamma_1, \gamma_2)\boldsymbol{c}_1, \qquad (15)$$

where $M(\gamma_1, \gamma_2)$ is defined in Lemma 6. Note that the definition of $M(\gamma_1, \gamma_2)$ here corresponds precisely to the definition of $\boldsymbol{M}(\gamma_1, \gamma_2)$ in Lemma 1 if $\boldsymbol{A}_1, \boldsymbol{A}_2$ satisfy the block-wise orthogonality assumption. Moreover, A_1 and A_2 already satisfy this assumption, therefore we can now make the following conclusion by invoking Theorem 2.

Corollary 1. *Assume* A *has full column rank and the* $n_2 \times n_1$ *matrix* $C := A_2^\top A_1$ *is nonzero. The following hold.*

- *If* C *has full rank (i.e.,* $\mathrm{rank}(C) = \min\{n_1, n_2\}$*), then*

$$\min_{\gamma_1 > 0, \gamma_2 > 0} \rho\big(M(\gamma_1, \gamma_2)\big) = \frac{\sin(\theta_r) - \sin(\theta_1)}{\sin(\theta_r) + \sin(\theta_1)}.$$

- *If* C *is rank-deficient, then*

$$\min_{\gamma_1 > 0, \gamma_2 > 0} \rho\big(M(\gamma_1, \gamma_2)\big) = \frac{1 - \sin(\theta_1)}{1 + \sin(\theta_1)}.$$

Note that for two *generic* subspaces $\mathcal{H}_1, \mathcal{H}_2$, with probability 1 the following hold: The matricx $A = [A_1, \ A_2]$ has full column rank if $n_1 + n_2 \leq m$; and $A_2^\top A_1$ has full rank.

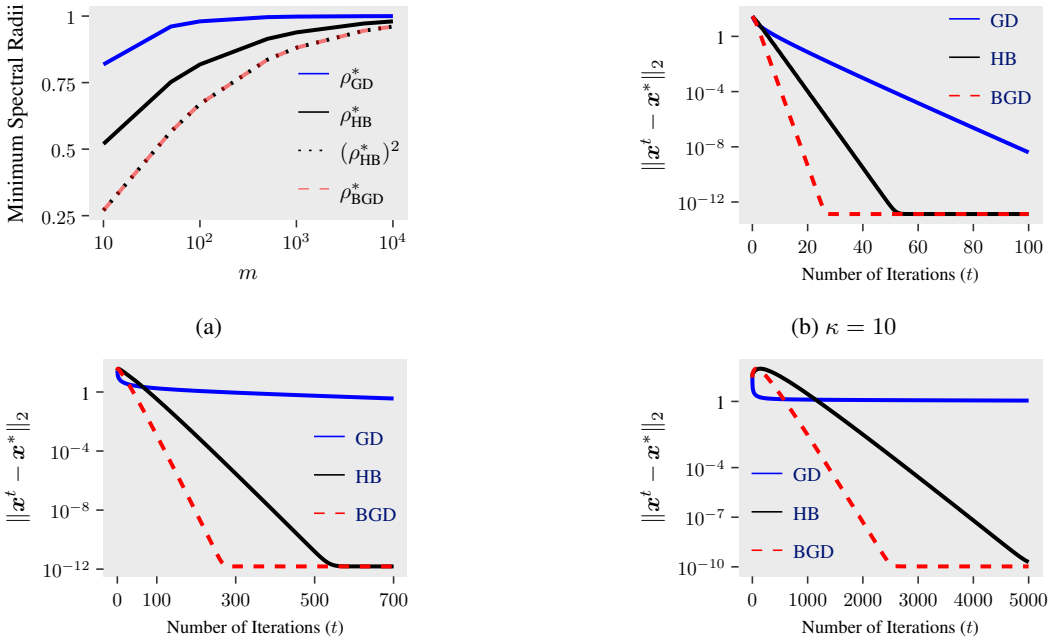
7. Numerical Experiments

In this section, we perform a simple experiment in order to validate our theory. The full MATLAB code for the experiment can be found in Appendix F.1.

Setup. We generate the linear regression data $\boldsymbol{A} = [\boldsymbol{A}_1 \ \boldsymbol{A}_2]$ and $\boldsymbol{y}$ randomly, via the function `gen_data` (Appendix F), such that Assumption 1 is fulfilled and the condition number of $\boldsymbol{A}$ can be specified as an input of the function. We implement BGD with stepsizes in (7), and we implement GD and HB with their optimal stepsizes. Note that Assumption 1 puts no restrictions on the spectrum of $\boldsymbol{A}$, hence, even under Assumption 1, the optimal stepsizes of GD and HB remain to be the same as reviewed in Section 2.1.

Results. In Figure 1a, we plot under Assumption 1 the numerical values of $\rho^*_{\mathrm{GD}}, \rho^*_{\mathrm{HB}}$, and ρ^*_{BGD}. In Figures 1b to 1d, we plot the distances $\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$ of the iterates $\{\boldsymbol{x}^t\}_t$ of GD, HB, and BGD for different condition numbers $\kappa := \frac{\lambda_1(\boldsymbol{A}^\top \boldsymbol{A})}{\lambda_n(\boldsymbol{A}^\top \boldsymbol{A})}$. Observe that the numerical convergence rates of HB and BGD are following our theory and especially the inequality $\rho^*_{\mathrm{BGD}} \leq (\rho^*_{\mathrm{HB}})^2$ (Remark 2).

Discussion. Can the proposed stepsizes for BGD be applied to general least-squares problems where $\boldsymbol{A} = [\boldsymbol{A}_1 \ \boldsymbol{A}_2]$ do not satisfy Assumption 1? As already indicated in Section 3.1, the answer is affirmative, and the idea is that one could always orthogonalize $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, respectively, e.g., via QR decomposition $\boldsymbol{A}_1 = \boldsymbol{Q}_1 \boldsymbol{R}_1$ and $\boldsymbol{A}_2 = \boldsymbol{Q}_2 \boldsymbol{R}_2$. Indeed, one could then consider minimizing $\|[\boldsymbol{Q}_1 \ \boldsymbol{Q}_2]\boldsymbol{z} - \boldsymbol{y}\|_2$ in variable $\boldsymbol{z}$ via BGD with the proposed stepsizes as the matrix $[\boldsymbol{Q}_1 \ \boldsymbol{Q}_2]$ now satisfies Assumption 1. After such minimization, we obtain minimizer $\boldsymbol{z}$ and just need to solve the upper triangular systems $\boldsymbol{z}_j = \boldsymbol{R}_j \boldsymbol{x}_j$ for $\boldsymbol{x}_j$ ($j = 1, 2$) by efficient back substitution. But how does this algorithm compare to HB applied directly to the data $\boldsymbol{A}, \boldsymbol{y}$? Note that $[\boldsymbol{Q}_1 \ \boldsymbol{Q}_2]$ is, in general, much better conditioned than $\boldsymbol{A}$ and its block-wise orthogonality structure allows for faster matrix-vector multiplications $[\boldsymbol{Q}_1 \ \boldsymbol{Q}_2]^\top [\boldsymbol{Q}_1 \ \boldsymbol{Q}_2]\boldsymbol{z}$ (e.g., already we have $\boldsymbol{Q}_1^\top \boldsymbol{Q}_1 \boldsymbol{z}_1 = \boldsymbol{z}_1$) than calculating $\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x}$.



(a)

(b) $\kappa = 10$

(c) $\kappa = 10^3$

(d) $\kappa = 10^5$

Figure 1: Under Assumption 1, Figure 1a shows the numerical values of the minimum spectral radii, ρ^*_{GD}, ρ^*_{HB}, and ρ^*_{BGD}, and Figures 1b to 1d shows the errors of the three methods at every iteration t.

For these two reasons, we observe that combining block-wise QR orthogonalization with BGD takes much fewer iterations than HB to converge and uses less computation at each iteration; and eventually this method runs faster than HB when $\boldsymbol{A}$ is highly ill-conditioned (so that HB converges slowly) and for medium problem sizes (so that block-wise QR decompositions are affordable). The code and experiments comparing the two approaches are in Appendix F.2.

8. Conclusion and Future Work

In this paper, we analyzed a basic algorithm, block gradient descent (BGD), applied to a basic setting, least-squares. Under a block-wise orthogonality assumption, we discover optimal stepsizes, with which BGD accelerates without any momentum and is provably twice as fast as the heavy ball method (HB). Moreover, our results apply to (generalized) alternating projection; Doing so not only leads to improvements over the bounds of Fält & Giselsson (2017), but also allows us to revise their inaccurate claim that an optimal convergence rate for (generalized) alternating projection was discovered. Among many interesting questions that might follow from our developments in this paper, we would like to elaborate on the following three points:

• Does adding some momentum term to BGD give faster rates? Goujaud et al. (2022) showed that GD with cyclic stepsizes and HB momentum is faster than HB if the spectrum of $\boldsymbol{A}^\top \boldsymbol{A}$ is clustered, so the answer might lie in combining the contributions of Goujaud et al. (2022) and ours.

• Can we generalize our analysis from two blocks to multiple blocks? If BGD with two blocks can run twice as fast as HB, would it be natural—or bold—to guess that BGD with n blocks can be n times faster? Moreover, in the case of n blocks, our block-wise orthogonality assumption (Assumption 1) would generalize to the less restrictive requirement that the columns of $\boldsymbol{A}$ are normalized; concerns regarding Assumption 1 would automatically disappear.

• To what extent can we dispense with Assumption 1? Recall that, without this assumption, we need to analyze the spectral radius of the sophisticated matrix $\boldsymbol{M}(\gamma_1, \gamma_2)$ defined in Lemma 1. While in this general case it is significantly more challenging to find a closed-form solution to the minimum spectral radius of $\boldsymbol{M}(\gamma_1, \gamma_2)$, it is perhaps possible to compute a numerical or approximate solution to (3), e.g., via semidefinite relaxation techniques. Pushing this idea to its extreme constitutes an entirely different chapter, and is therefore left as a future endeavor.

Impact Statement

This paper presents work intending to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agarwal, N., Goel, S., and Zhang, C. Acceleration via fractal learning rate schedules. In *International Conference on Machine Learning*, pp. 87–99. PMLR, 2021.

Altschuler, J. M. and Parrilo, P. A. Acceleration by stepsize hedging i: Multi-step descent and the silver stepsize schedule. Technical report, arXiv:2309.07879 [math.OC], 2023a.

Altschuler, J. M. and Parrilo, P. A. Acceleration by stepsize hedging ii: Silver stepsize schedule for smooth convex optimization. Technical report, arXiv:2309.16530 [math.OC], 2023b.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Bauschke, H. H., Cruz, J. B., Nghia, T. T., Phan, H. M., and Wang, X. 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.

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

Beck, A. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017.

Beck, A. and Tetruashvili, L. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

Cai, X. and Diakonikolas, J. Last iterate convergence of incremental methods and applications in continual learning. Technical report, arXiv:2403.06873 [math.OC], 2024.

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

Deutsch, F. Rate of convergence of the method of alternating projections. *Parametric Optimization and Approximation*, 72:96–107, 1984.

Douglas, J. and Rachford, H. H. 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.

Elenter, J., NaderiAlizadeh, N., Javidi, T., and Ribeiro, A. Primal-dual continual learning: Stability and plasticity through lagrange multipliers. Technical report, arXiv:2310.00154 [cs.LG], 2023.

Evron, I., Moroshko, E., Ward, R., Srebro, N., and Soudry, D. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, 2022.

Fält, M. and Giselsson, P. Optimal convergence rates for generalized alternating projections. In *IEEE Annual Conference on Decision and Control*, 2017.

Franchetti, C. and Light, W. On the Von Neumann alternating algorithm in Hilbert space. *Journal of Mathematical Analysis and Applications*, 114(2):305–314, 1986.

Friedrichs, K. On certain inequalities and characteristic value problems for analytic functions and for functions of two variables. *Transactions of the American Mathematical Society*, 41(3):321–364, 1937.

Goujaud, B., Scieur, D., Dieuleveut, A., Taylor, A. B., and Pedregosa, F. Super-acceleration with cyclical step-sizes. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Grimmer, B. Provably faster gradient descent via long steps. Technical report, arXiv:2307.06324 [math.OC], 2023.

Hesse, R., Luke, D. R., and Neumann, P. Alternating projections and Douglas-Rachford for sparse affine feasibility. *IEEE Transactions on Signal Processing*, 62(18):4868–4881, 2014.

Hildreth, C. A quadratic programming procedure. *Naval Research Logistics Quarterly*, 4(1):79–85, 1957.

Kayalar, S. and Weinert, H. L. Error bounds for the method of alternating projections. *Mathematics of Control, Signals and Systems*, 1(1):43–59, 1988.

Knyazev, A. V. and Argentati, M. E. Majorization for changes in angles between subspaces, Ritz values, and graph Laplacian spectra. *SIAM Journal on Matrix Analysis and Applications*, 29(1):15–32, 2007.

Lin, C. Y., Song, C., and Diakonikolas, J. Accelerated cyclic coordinate dual averaging with extrapolation for composite convex optimization. In *International Conference on Machine Learning*, 2023.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2018a.

Nesterov, Y. *Lectures on Convex Optimization*. Springer, 2018b.

Nutini, J., Schmidt, M., Laradji, I., Friedlander, M., and Koepke, H. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *International Conference on Machine Learning*, 2015.

Oymak, S. Provable super-convergence with a large cyclical learning rate. *IEEE Signal Processing Letters*, 28:1645–1649, 2021.

Peng, L. and Vidal, R. Block coordinate descent on smooth manifolds. Technical report, arXiv:2305.14744 [math.OC], 2023.

Peng, L., Giampouras, P., and Vidal, R. The ideal continual learner: An agent that never forgets. In *International Conference on Machine Learning*, 2023.

Peng, Z., Wu, T., Xu, Y., Yan, M., and Yin, W. Coordinate-friendly structures, algorithms and applications. *Annals of Mathematical Sciences and Applications*, 1(1):57–119, 2016.

Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Polyak, B. T. *Introduction to Optimization*. Optimization Software, 1987.

Shi, H.-J. M., Tu, S., Xu, Y., and Yin, W. A primer on coordinate descent algorithms. Technical report, arXiv:1610.00040v2 [math.OC], 2016.

Smith, L. N. Cyclical learning rates for training neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2017.

Sun, R. and Ye, Y. Worst-case complexity of cyclic coordinate descent: $o(n^2)$ gap with randomized version. *Mathematical Programming*, 185:487–520, 2021.

Vidal, R., Ma, Y., and Sastry, S. *Generalized Principal Component Analysis*. Springer, 2016.

von Neumann, J. *Functional Operators (AM-22), Volume 2*. Princeton University Press, Princeton, 1951.

Wright, S. J. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

Xu, Y. and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

Young, D. On Richardson's method for solving linear systems with positive definite matrices. *Journal of Mathematics and Physics*, 32(1-4):243–255, 1953.

Zhu, P. and Knyazev, A. V. Angles between subspaces and their tangents. *Journal of Numerical Mathematics*, 21(4): 325–340, 2013.

A. Elementary Lemmas

Here we first give proofs to Lemmas 1 to 4:

Proof of Lemma 1. Note that the global minimizer $\boldsymbol{x}^*$ of (2) satisfies the normal equation $\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}^\top \boldsymbol{y}$, which means $\boldsymbol{A}_1^\top \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}_1^\top \boldsymbol{y}$ and $\boldsymbol{A}_2^\top \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}_2^\top \boldsymbol{y}$, so we can write the first equation $\boldsymbol{x}_1^+ = \boldsymbol{x}_1 - \gamma_1 \cdot \boldsymbol{A}_1^\top \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*)$ of BGD as

$$\begin{bmatrix} \boldsymbol{x}_1^+ \\ \boldsymbol{x}_2 \end{bmatrix} - \boldsymbol{x}^* = \left(\boldsymbol{I} - \begin{bmatrix} \gamma_1 \boldsymbol{A}_1^\top \boldsymbol{A}_1 & \gamma_1 \boldsymbol{A}_1^\top \boldsymbol{A}_2 \\ 0 & 0 \end{bmatrix} \right) (\boldsymbol{x} - \boldsymbol{x}^*).$$

We finish by re-writing the second equation of BGD similarly and combining. □

Proof of Lemma 2. Left multiplying the first equation of BGD by $\boldsymbol{A}_1$ and the second equation by $\boldsymbol{A}_2$ gives

$$\begin{aligned} \boldsymbol{A}_1 \boldsymbol{x}_1^+ &= \boldsymbol{A}_1 \boldsymbol{x}_1 - \gamma_1 \cdot \boldsymbol{A}_1 \boldsymbol{A}_1^\top (\boldsymbol{A}_1 \boldsymbol{x}_1 + \boldsymbol{A}_2 \boldsymbol{x}_2 - \boldsymbol{y}), \\ \boldsymbol{A}_2 \boldsymbol{x}_2^+ &= \boldsymbol{A}_2 \boldsymbol{x}_2 - \gamma_2 \cdot \boldsymbol{A}_2 \boldsymbol{A}_2^\top (\boldsymbol{A}_1 \boldsymbol{x}_1^+ + \boldsymbol{A}_2 \boldsymbol{x}_2 - \boldsymbol{y}). \end{aligned}$$

Substitute the expression of $\boldsymbol{A}_1 \boldsymbol{x}_1^+$ into the second, and sum them up, and we obtain

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x}^+ &= \boldsymbol{A}\boldsymbol{x} - \gamma_1 \boldsymbol{A}_1 \boldsymbol{A}_1^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) - \gamma_2 \cdot \boldsymbol{A}_2 \boldsymbol{A}_2^\top \big(\boldsymbol{A}_1 \boldsymbol{x}_1 - \gamma_1 \cdot \boldsymbol{A}_1 \boldsymbol{A}_1^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) + \boldsymbol{A}_2 \boldsymbol{x}_2 - \boldsymbol{y} \big) \\ &= \boldsymbol{A}\boldsymbol{x} - \gamma_1 \boldsymbol{A}_1 \boldsymbol{A}_1^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) - \gamma_2 \cdot \boldsymbol{A}_2 \boldsymbol{A}_2^\top \big(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} - \gamma_1 \cdot \boldsymbol{A}_1 \boldsymbol{A}_1^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) \big) \\ &= \boldsymbol{A}\boldsymbol{x} - \gamma_1 \boldsymbol{A}_1 \boldsymbol{A}_1^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) - \gamma_2 \cdot \boldsymbol{A}_2 \boldsymbol{A}_2^\top \big(\boldsymbol{I} - \gamma_1 \cdot \boldsymbol{A}_1 \boldsymbol{A}_1^\top \big) (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}). \end{aligned}$$

As shown in the proof of Lemma 1 we have $\boldsymbol{A}_1^\top \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}_1^\top \boldsymbol{y}$ and $\boldsymbol{A}_2^\top \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}_2^\top \boldsymbol{y}$. Subtracting $\boldsymbol{A}\boldsymbol{x}^*$ the above equation yields

$$\boldsymbol{A}(\boldsymbol{x}^+ - \boldsymbol{x}^*) = \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*) - \gamma_1 \boldsymbol{A}_1 \boldsymbol{A}_1^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}^*) - \gamma_2 \cdot \boldsymbol{A}_2 \boldsymbol{A}_2^\top \big(\boldsymbol{I} - \gamma_1 \cdot \boldsymbol{A}_1 \boldsymbol{A}_1^\top \big) (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}^*).$$

The proof is finished by simplifying the above equation. □

Proof of Lemma 3. Under Assumption 1 and with $\boldsymbol{C} := \boldsymbol{A}_2^\top \boldsymbol{A}_1$, we have

$$\begin{aligned} \boldsymbol{M}(\gamma_1, \gamma_2) &= \left(\boldsymbol{I} - \begin{bmatrix} 0 & 0 \\ \gamma_2 \boldsymbol{A}_2^\top \boldsymbol{A}_1 & \gamma_2 \boldsymbol{A}_2^\top \boldsymbol{A}_2 \end{bmatrix} \right) \left(\boldsymbol{I} - \begin{bmatrix} \gamma_1 \boldsymbol{A}_1^\top \boldsymbol{A}_1 & \gamma_1 \boldsymbol{A}_1^\top \boldsymbol{A}_2 \\ 0 & 0 \end{bmatrix} \right) \\ &= \left(\boldsymbol{I} - \begin{bmatrix} 0 & 0 \\ \gamma_2 \boldsymbol{C} & \gamma_2 \boldsymbol{I}_2 \end{bmatrix} \right) \left(\boldsymbol{I} - \begin{bmatrix} \gamma_1 \boldsymbol{I}_1 & \gamma_1 \boldsymbol{C}^\top \\ 0 & 0 \end{bmatrix} \right) \\ &= \boldsymbol{I} - \begin{bmatrix} 0 & 0 \\ \gamma_2 \boldsymbol{C} & \gamma_2 \boldsymbol{I}_2 \end{bmatrix} - \begin{bmatrix} \gamma_1 \boldsymbol{I}_1 & \gamma_1 \boldsymbol{C}^\top \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \gamma_2 \boldsymbol{C} & \gamma_2 \boldsymbol{I}_2 \end{bmatrix} \begin{bmatrix} \gamma_1 \boldsymbol{I}_1 & \gamma_1 \boldsymbol{C}^\top \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} (1-\gamma_1)\boldsymbol{I}_1 & -\gamma_1 \boldsymbol{C}^\top \\ -\gamma_2 \boldsymbol{C} & (1-\gamma_2)\boldsymbol{I}_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \gamma_1\gamma_2 \boldsymbol{C} & \gamma_1\gamma_2 \boldsymbol{C}\boldsymbol{C}^\top \end{bmatrix} \\ &= \begin{bmatrix} (1-\gamma_1)\boldsymbol{I}_1 & -\gamma_1 \boldsymbol{C}^\top \\ -\gamma_2(1-\gamma_1)\boldsymbol{C} & (1-\gamma_2)\boldsymbol{I}_2 + \gamma_1\gamma_2 \boldsymbol{C}\boldsymbol{C}^\top \end{bmatrix}, \end{aligned}$$

and the proof is complete. □

Proof of Lemma 4. By Assumption 1 we write $\boldsymbol{A}^\top \boldsymbol{A}$ as

$$\boldsymbol{A}^\top \boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_1 & \boldsymbol{C}^\top \\ \boldsymbol{C} & \boldsymbol{I}_2 \end{bmatrix} = \boldsymbol{I} + \boldsymbol{D}, \quad \boldsymbol{D} := \begin{bmatrix} 0 & \boldsymbol{C}^\top \\ \boldsymbol{C} & 0 \end{bmatrix}.$$

We will first find the eigenvalues of the $(n_1 + n_2) \times (n_1 + n_2)$ matrix $\boldsymbol{D}$. Since $\boldsymbol{C}$ has rank r, $\boldsymbol{D}$ has eigenvalue 0 of geometric multiplicity $n_1 + n_2 - 2r$, and so the algebraic multiplicity of eigenvalue 0 is at least $n_1 + n_2 - 2r$. With a variable $\lambda \neq 0$ we now look at the characteristic polynomial $\det(\lambda \boldsymbol{I} - \boldsymbol{D})$. By Lemma 7 we have

$$\det(\lambda \boldsymbol{I} - \boldsymbol{D}) = \det(\lambda \boldsymbol{I}_1) \cdot \det\left(\lambda \boldsymbol{I}_2 - \frac{1}{\lambda} \boldsymbol{C}\boldsymbol{C}^\top \right).$$

Therefore, $\det(\lambda \boldsymbol{I} - \boldsymbol{D}) = 0$ if and only if $\det\left(\lambda^2 \boldsymbol{I}_2 - \boldsymbol{C}\boldsymbol{C}^\top\right) = 0$. Since $\lambda \neq 0$, we have $\det\left(\lambda^2 \boldsymbol{I}_2 - \boldsymbol{C}\boldsymbol{C}^\top\right) = 0$ if and only if λ^2 is a positive eigenvalue of $\boldsymbol{C}\boldsymbol{C}^\top$. Since $\boldsymbol{C}\boldsymbol{C}^\top$ has r such eigenvalues $\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top), \ldots, \lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)$, we have $2r$ such choices for λ, namely $\pm\sqrt{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}, \ldots, \pm\sqrt{\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)}$. Eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$ are eigenvalues of $\boldsymbol{D}$ plus 1. We finish by recalling Section 2 and simple calculation. □

Lemmas 7 to 12 below are elementary, so their proofs are omitted.

Lemma 7 (Block Matrix Determinant)**.** *We have*

$$\det \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \\ \boldsymbol{Q}_3 & \boldsymbol{Q}_4 \end{bmatrix} = \det(\boldsymbol{Q}_1) \cdot \det(\boldsymbol{Q}_4 - \boldsymbol{Q}_3 \boldsymbol{Q}_1^{-1} \boldsymbol{Q}_2),$$

where $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_4$ are square matrices with $\boldsymbol{Q}_1$ invertible.

Lemma 8. *Let $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_r > 0$. Then we have*

$$\min_{\gamma > 0} \Big(\max\left\{ |1 - \gamma \xi_1|, \ldots, |1 - \gamma \xi_r| \right\} \Big) = \begin{cases} \frac{\xi_1 - \xi_r}{\xi_1 + \xi_r} & r > 1; \\ \frac{1}{\xi_1} & r = 1. \end{cases}$$

Here the minimum is attained at $\gamma = \frac{2}{\xi_1 + \xi_r}$ if $r > 1$ or at $\gamma = \frac{1}{\xi_1}$ if $r = 1$.

Lemma 9. *For functions $f_i : S \to \mathbb{R}$ ($\forall i = 1, \ldots, r$), we have*

$$\hat{\gamma} \in \operatorname*{argmin}_{\gamma \in S} f_i(\gamma), \forall i = 1, \ldots, r \Rightarrow \hat{\gamma} \in \operatorname*{argmin}_{\gamma \in S} \Big(\max_{i=1,\ldots,r} f_i(\gamma) \Big).$$

Lemma 10. *With $\phi \in (0, 1)$, $\gamma_1 \in (0, 1)$, and $\hat{\gamma}_2 := \frac{4 - 2\gamma_1}{2 - \gamma_1 \phi}$, the quadratic equation*

$$z^2 - (2 - \gamma_1 - \hat{\gamma}_2 + \gamma_1 \hat{\gamma}_2 \phi) z + (1 - \gamma_1)(1 - \hat{\gamma}_2) = 0$$

in variable z has two real roots, $z = \gamma_1 - 1$ and $z = \hat{\gamma}_2 - 1$ with $\hat{\gamma}_2 - 1 > 1 - \gamma_1$.

Lemma 11. *For functions $f_i : S \to \mathbb{R}$ ($\forall i = 0, 1, \ldots, r$), if*

$$\hat{\gamma} \in \operatorname*{argmin}_{\gamma \in S} \Big(\max\{ f_0(\gamma), f_1(\gamma) \} \Big)$$

and $f_i(\hat{\gamma}) \leq \max\{f_0(\hat{\gamma}), f_1(\hat{\gamma})\}$ for every $i = 1, \ldots, r$, then

$$\hat{\gamma} \in \operatorname*{argmin}_{\gamma \in S} \Big(\max_{i=0,\ldots,r} f_i(\gamma) \Big).$$

Lemma 12. *Consider the two (perhaps complex) roots z_1 and z_2 to the quadratic equation*

$$z^2 + bz + c = 0$$

in variable z with $c \geq 0$ and $b \in \mathbb{R}$. We have

$$|z_1| \leq \sqrt{c} \text{ and } |z_2| \leq \sqrt{c} \Leftrightarrow |z_1| = |z_2| = \sqrt{c} \Leftrightarrow b^2 - 4c \leq 0.$$

B. Proof of The Main Result (Theorem 2)

B.1. A Spectrum Lemma

The first step towards solving (3) is to analyze the spectrum of $\boldsymbol{M}(\gamma_1, \gamma_2)$. We do so in the lemma below.

Lemma 13 (Spectrum of $\boldsymbol{M}(\gamma_1, \gamma_2)$)**.** *Suppose Assumption 1 holds and $\boldsymbol{C} = \boldsymbol{A}_2^\top \boldsymbol{A}_1 \in \mathbb{R}^{n_2 \times n_1}$ has rank r. Then, besides the eigenvalues $1 - \gamma_1$ and $1 - \gamma_2$ shown in Table 2, the remaining eigenvalues of $\boldsymbol{M}(\gamma_1, \gamma_2)$ are given as follows:*

- *(Case 1: $\gamma_1 = 1$) The remaining r eigenvalues of $\boldsymbol{M}(\gamma_1, \gamma_2)$ are $1 - \gamma_2 + \gamma_2 \lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)$, $\forall i = 1, \ldots, r$;*
- *(Case 2: $\gamma_2 = 1$) The remaining r eigenvalues of $\boldsymbol{M}(\gamma_1, \gamma_2)$ are $1 - \gamma_1 + \gamma_1 \lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)$, $\forall i = 1, \ldots, r$;*

Table 2: Algebraic multiplicities of eigenvalues $1-\gamma_1$ and $1-\gamma_2$ of $\boldsymbol{M}(\gamma_1,\gamma_2)$. The rank of $\boldsymbol{C}$ is denoted by r. The table accompanies Lemma 13.

		Eigenvalues	
		$1-\gamma_1$	$1-\gamma_2$
Cases	$\gamma_1=1$	n_1	n_2-r
	$\gamma_2=1$	n_1-r	n_2
	$\gamma_1\neq 1,\gamma_2\neq 1$	n_1-r	n_2-r

• *(Case 3: $\gamma_1\neq 1,\gamma_2\neq 1$) The remaining $2r$ eigenvalues of $\boldsymbol{M}(\gamma_1,\gamma_2)$ arise as the roots of the following quadratic equation in variable z ($\forall i=1,\dots,r$):*

$$z^2-\big(2-\gamma_1-\gamma_2+\gamma_1\gamma_2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\big)z+(1-\gamma_1)(1-\gamma_2)=0. \tag{16}$$

Remark 4. The proof of Lemma 13 is lengthy, and the difficulty lies in the fact that stepsizes γ_1,γ_2 could change the eigenvalues of $\boldsymbol{M}(\gamma_1,\gamma_2)$ in a "discontinuous" way (cf. Table 2). And we need to handle all such discontinuities.

Remark 5. Lemma 13, in conjunction with Table 2, characterizes the spectrum of $\boldsymbol{M}(\gamma_1,\gamma_2)$ in different cases. Case 1 ($\gamma_1=1$) and Case 2 ($\gamma_2=1$) turn out to be symmetric, and they coincide if $\gamma_1=\gamma_2=1$. In either case, the eigenvalues of $\boldsymbol{M}(\gamma_1,\gamma_2)$ are given explicitly, in terms of $\gamma_1,\gamma_2,\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)$. In Case 3 ($\gamma_1\neq 1,\gamma_2\neq 1$), we have $2r$ eigenvalues implicitly defined by quadratic equations (16).

Proof of Lemma 13. Under Assumption 1, $\boldsymbol{M}(\gamma_1,\gamma_2)$ is of the form in Lemma 3. We approach the proof by considering the three cases corresponding to Table 2 separately.

Case 1: $\gamma_1=1$. In this case, $1-\gamma_1=0$ and $\boldsymbol{M}(\gamma_1,\gamma_2)$ is given as

$$\boldsymbol{M}(1,\gamma_2)=\begin{bmatrix}0 & -\boldsymbol{C}^\top\\ 0 & (1-\gamma_2)\boldsymbol{I}_2+\gamma_2\boldsymbol{C}\boldsymbol{C}^\top\end{bmatrix}.$$

The eigenvalues of $(1-\gamma_2)\boldsymbol{I}_2+\gamma_2\boldsymbol{C}\boldsymbol{C}^\top$ give n_2 eigenvalues to $\boldsymbol{M}(1,\gamma_2)$; they are $1-\gamma_2$ and $(1-\gamma_2)+\gamma_2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)$, $i=1,\dots,r$. Note that $1-\gamma_2$ has geometric multiplicity n_2-r matching the dimension of the nullspace of $\boldsymbol{C}\boldsymbol{C}^\top$. Finally, 0 is an eigenvalue of $\boldsymbol{M}(1,\gamma_2)$ with geometric multiplicity n_1. Since all n_1+n_2 eigenvalues are counted. the abovementioned geometric multiplicity coincides with algebraic multiplicity. Similarly, in the rest of the proof, the eigenvalues will be counted using either geometric or algebraic ways, but by finding all eigenvalues (counting geometric or algebraic multiplicity) we will eventually reveal that in our case the geometric multiplicity turns out to coincide with algebraic multiplicity. Hence, in what follows we will just write *multiplicity* for simplicity.

Case 2: $\gamma_2=1$. In this case, $1-\gamma_2=0$ and $\boldsymbol{M}(\gamma_1,\gamma_2)$ is given as

$$\boldsymbol{M}(\gamma_1,1)=\begin{bmatrix}(1-\gamma_1)\boldsymbol{I}_1 & -\gamma_1\boldsymbol{C}^\top\\ -(1-\gamma_1)\boldsymbol{C} & \gamma_1\boldsymbol{C}\boldsymbol{C}^\top\end{bmatrix}.$$

If $\gamma_1=1$, by *Case 1* we know that $0=1-\gamma_1=1-\gamma_2$ is an eigenvalue of $\boldsymbol{M}(\gamma_1,1)$ with multiplicity n_1+n_2-r, and the remaining r eigenvalues are given as $\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)$, $i=1,\dots,r$; Table 2 is correct. So in the remaining proof of *Case 2* we assume $\gamma_1\neq 1$.

Case 2.1: Eigenvalue $1-\gamma_1$. Let us now test whether $\boldsymbol{M}(\gamma_1,1)$ has eigenvalue $1-\gamma_1$. With an (n_1+n_2)-dimensional vector $\boldsymbol{v}:=[\boldsymbol{v}_1;\boldsymbol{v}_2]$ we have

$$\begin{aligned}\boldsymbol{M}(\gamma_1,1)\boldsymbol{v}=(1-\gamma_1)\boldsymbol{v}&\Leftrightarrow\begin{cases}(1-\gamma_1)\boldsymbol{v}_1-\gamma_1\boldsymbol{C}^\top\boldsymbol{v}_2=(1-\gamma_1)\boldsymbol{v}_1\\ -(1-\gamma_1)\boldsymbol{C}\boldsymbol{v}_1+\gamma_1\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{v}_2=(1-\gamma_1)\boldsymbol{v}_2\end{cases}\\ &\Leftrightarrow\begin{cases}\boldsymbol{C}^\top\boldsymbol{v}_2=0\\ \boldsymbol{C}\boldsymbol{v}_1=-\boldsymbol{v}_2\end{cases}\end{aligned}$$

and this implies $\boldsymbol{C}^\top \boldsymbol{C}\boldsymbol{v}_1 = -\boldsymbol{C}^\top \boldsymbol{v}_2 = 0$. Hence $\boldsymbol{v}_2 = \boldsymbol{C}\boldsymbol{v}_1 = 0$. On the other hand, if $\boldsymbol{C}\boldsymbol{v}_1 = \boldsymbol{v}_2 = 0$ then the above is satisfied. We now conclude that $\boldsymbol{M}(\gamma_1, 1)\boldsymbol{v} = (1-\gamma_1)\boldsymbol{v}$ if and only if $\boldsymbol{C}\boldsymbol{v}_1 = \boldsymbol{v}_2 = 0$. The nullspace of $\boldsymbol{C}$ is of dimension $n_1 - r$, so $(1-\gamma_1)$ is an eigenvalue of $\boldsymbol{M}(\gamma_1, 1)$ of multiplicity $n_1 - r$.

Case 2.2: Eigenvalue 0. Following a similar proof of *Case 2.1*, we have

$$\begin{aligned}\boldsymbol{M}(\gamma_1,1)\boldsymbol{v} = 0 &\Leftrightarrow \begin{cases}(1-\gamma_1)\boldsymbol{v}_1 - \gamma_1 \boldsymbol{C}^\top \boldsymbol{v}_2 = 0\\ -(1-\gamma_1)\boldsymbol{C}\boldsymbol{v}_1 + \gamma_1 \boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{v}_2 = 0\end{cases}\\ &\Leftrightarrow (1-\gamma_1)\boldsymbol{v}_1 - \gamma_1 \boldsymbol{C}^\top \boldsymbol{v}_2 = 0.\end{aligned}$$

Since $\gamma_1 \neq 1$, for any $\boldsymbol{v}_2 \in \mathbb{R}^{n_2}$ we can set $\boldsymbol{v}_1$ to be $\gamma_1 \boldsymbol{C}^\top \boldsymbol{v}_2/(1-\gamma_1)$ so that the equality $\boldsymbol{M}(\gamma_1, 1)\boldsymbol{v} = 0$ holds. Therefore 0 is an eigenvalue of $\boldsymbol{M}(\gamma_1, 1)$ of multiplicity n_2.

Case 2.3: The Remaining r Eigenvalues. Consider the characteristic polynomial $\det\big(z\boldsymbol{I} - \boldsymbol{M}(\gamma_1, 1)\big)$ in variable z, with $z \neq 1-\gamma_1$ and $z \neq 0$. Lemma 7 implies

$$\begin{aligned}\det\big(z\boldsymbol{I} - \boldsymbol{M}(\gamma_1,1)\big) = \det\Big(z\boldsymbol{I}_1 - (1-\gamma_1)\boldsymbol{I}_1\Big)\cdot \det\Big(z\boldsymbol{I}_2 - \gamma_1 \boldsymbol{C}\boldsymbol{C}^\top \\ - (1-\gamma_1)\boldsymbol{C}\cdot \frac{1}{z-(1-\gamma_1)}\cdot \gamma_1 \boldsymbol{C}^\top\Big).\end{aligned}$$

With $z \neq 1-\gamma_1$ and $z \neq 0$, the above implies

$$\det\big(z\boldsymbol{I} - \boldsymbol{M}(\gamma_1,1)\big) = 0 \Leftrightarrow \det\Big(z\boldsymbol{I}_2 - (1-\gamma_1)\boldsymbol{I}_2 - \gamma_1 \boldsymbol{C}\boldsymbol{C}^\top\Big) = 0. \tag{17}$$

We now see that for each $i = 1, \ldots, r$, $(1-\gamma_1) + \gamma_1 \lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)$ is a solution to (17), and is therefore an eigenvalue of $\boldsymbol{M}(\gamma_1, 1)$.

Case 3: $\gamma_1 \neq 1, \gamma_2 \neq 1$. Suppose for the sake of contradiction that $\boldsymbol{M}(\gamma_1, \gamma_2)\boldsymbol{v} = 0$ for some $\boldsymbol{v} = [\boldsymbol{v}_1; \boldsymbol{v}_2] \in \mathbb{R}^{n_1+n_2}$. Then, using the formula of $\boldsymbol{M}(\gamma_1, \gamma_2)$ in Lemma 3 we can write

$$\begin{aligned}&\begin{cases}(1-\gamma_1)\boldsymbol{v}_1 - \gamma_1 \boldsymbol{C}^\top \boldsymbol{v}_2 = 0\\ -\gamma_2(1-\gamma_1)\boldsymbol{C}\boldsymbol{v}_1 + (1-\gamma_2)\boldsymbol{v}_2 + \gamma_1\gamma_2 \boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{v}_2 = 0\end{cases}\\ \Leftrightarrow &\begin{cases}(1-\gamma_1)\boldsymbol{v}_1 - \gamma_1 \boldsymbol{C}^\top \boldsymbol{v}_2 = 0\\ -\gamma_2(1-\gamma_1)\boldsymbol{C}\boldsymbol{v}_1 + (1-\gamma_2)\boldsymbol{v}_2 + \gamma_2 \boldsymbol{C}(1-\gamma_1)\boldsymbol{v}_1 = 0\end{cases}\\ \Leftrightarrow &\begin{cases}(1-\gamma_1)\boldsymbol{v}_1 + \gamma_1 \boldsymbol{C}^\top \boldsymbol{v}_2 = 0\\ (1-\gamma_2)\boldsymbol{v}_2 = 0\end{cases}\end{aligned}$$

and since $\gamma_1 \neq 1$ and $\gamma_2 \neq 1$, it is now clear that $\boldsymbol{v}_1 = \boldsymbol{v}_2 = 0$. So $\boldsymbol{M}(\gamma_1, \gamma_2)$ is full rank.

We will show there are $2r$ eigenvalues different than $1-\gamma_1$ and $1-\gamma_2$ and given by the roots of (16). Consider the characteristic polynomial $\det\big(z\boldsymbol{I} - \boldsymbol{M}(\gamma_1, \gamma_2)\big)$ in variable z, with $z \neq 1-\gamma_1$, $z \neq 1-\gamma_2$, and $z \neq 0$. By Lemma 7 we have

$$\begin{aligned}\det\big(z\boldsymbol{I} - \boldsymbol{M}(\gamma_1,\gamma_2)\big) = \det\Big(z\boldsymbol{I}_1 - (1-\gamma_1)\boldsymbol{I}_1\Big)\cdot \det\Big(z\boldsymbol{I}_2 - (1-\gamma_2)\boldsymbol{I}_2 - \gamma_1\gamma_2 \boldsymbol{C}\boldsymbol{C}^\top \\ - \gamma_2(1-\gamma_1)\boldsymbol{C}\frac{1}{z-(1-\gamma_1)}\cdot \gamma_1 \boldsymbol{C}^\top\Big).\end{aligned}$$

Simplifying the above equation yields

$$\det\big(z\boldsymbol{I} - \boldsymbol{M}(\gamma_1,\gamma_2)\big) = 0 \Leftrightarrow \det\Big(\frac{\big(z-(1-\gamma_2)\big)\cdot\big(z-(1-\gamma_1)\big)}{z}\boldsymbol{I}_2 - \gamma_1\gamma_2 \boldsymbol{C}\boldsymbol{C}^\top\Big) = 0.$$

The characteristic polynomial $\det(z'\boldsymbol{I}_2 - \gamma_1\gamma_2 \boldsymbol{C}\boldsymbol{C}^\top)$ in variable z' has r nonzero roots, namely $\gamma_1\gamma_2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top), \ldots, \gamma_1\gamma_2\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)$. Moreover, for each root z', the equation

$$\frac{\big(z-(1-\gamma_2)\big)\cdot\big(z-(1-\gamma_1)\big)}{z} = z' \Leftrightarrow z^2 - \big(2-\gamma_1-\gamma_2+z'\big)z + (1-\gamma_1)(1-\gamma_2) = 0$$

always has two nonzero (potentially complex) solutions, and there are $2r$ such solutions in total; none of these solutions is equal to $1-\gamma_1$ or $1-\gamma_2$. In particular, by construction, these $2r$ solutions must be eigenvalues of $\boldsymbol{M}(\gamma_1,\gamma_2)$.

It remains to show $1-\gamma_1$ and $1-\gamma_2$ are eigenvalues of $\boldsymbol{M}(\gamma_1,\gamma_2)$ with mulpliticity n_1-r and n_2-r, respectively; cf. Table 2. We consider two cases, $\gamma_1=\gamma_2$ and $\gamma_1\neq\gamma_2$.

Case 3.1: $\gamma_1\neq\gamma_2$**.** With some vector $\boldsymbol{v}:=[\boldsymbol{v}_1;\boldsymbol{v}_2]\in\mathbb{R}^{n_1+n_2}$ we have

$$\begin{aligned}\boldsymbol{M}(\gamma_1,\gamma_2)\boldsymbol{v}=(1-\gamma_1)\boldsymbol{v}&\Leftrightarrow\begin{cases}-\gamma_1\boldsymbol{C}^\top\boldsymbol{v}_2=0\\-\gamma_2(1-\gamma_1)\boldsymbol{C}\boldsymbol{v}_1+(\gamma_1-\gamma_2)\boldsymbol{v}_2+\gamma_1\gamma_2\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{v}_2=0\end{cases}\\&\Leftrightarrow\begin{cases}\boldsymbol{C}^\top\boldsymbol{v}_2=0;\\-\gamma_2(1-\gamma_1)\boldsymbol{C}\boldsymbol{v}_1+(\gamma_1-\gamma_2)\boldsymbol{v}_2=0.\end{cases}\end{aligned}$$

Left multiplying the last equation by $\boldsymbol{C}^\top$ yields $\gamma_2(1-\gamma_1)\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{v}_1=0$. Since $\gamma_1\neq1$ and $\gamma_2>0$, we obtain $\boldsymbol{C}\boldsymbol{v}_1=0$. Substituting it back gives $(\gamma_1-\gamma_2)\boldsymbol{v}_2=0$, but $\gamma_1\neq\gamma_2$, so it must be that $\boldsymbol{v}_2=0$. Therefore, for $\boldsymbol{M}(\gamma_1,\gamma_2)\boldsymbol{v}=(1-\gamma_1)\boldsymbol{v}$ to hold, it is necessary that $\boldsymbol{v}_2=0$ and $\boldsymbol{C}\boldsymbol{v}_1=0$; one verifies this is also sufficient, and thus

$$\boldsymbol{M}(\gamma_1,\gamma_2)\boldsymbol{v}=(1-\gamma_1)\boldsymbol{v}\Leftrightarrow\boldsymbol{C}\boldsymbol{v}_1=0,\boldsymbol{v}_2=0.$$

Since the nullspace of $\boldsymbol{C}$ has dimension n_1-r, so $1-\gamma_1$ is an eigenvalue of $\boldsymbol{M}(\gamma_1,\gamma_2)$ of multiplicity n_1-r. Proving $1-\gamma_2$ is an eigenvalue of $\boldsymbol{M}(\gamma_1,\gamma_2)$ of multiplicity n_2-r follows a similar route, detailed next. With $\boldsymbol{v}:=[\boldsymbol{v}_1;\boldsymbol{v}_2]\in\mathbb{R}^{n_1+n_2}$ we have

$$\begin{aligned}\boldsymbol{M}(\gamma_1,\gamma_2)\boldsymbol{v}=(1-\gamma_2)\boldsymbol{v}&\Leftrightarrow\begin{cases}-\gamma_1\boldsymbol{C}^\top\boldsymbol{v}_2=(\gamma_1-\gamma_2)\boldsymbol{v}_1\\-(1-\gamma_1)\boldsymbol{C}\boldsymbol{v}_1+\gamma_1\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{v}_2=0\end{cases}\\&\Leftrightarrow\begin{cases}-\gamma_1\boldsymbol{C}^\top\boldsymbol{v}_2=(\gamma_1-\gamma_2)\boldsymbol{v}_1\\-(1-\gamma_1)\boldsymbol{C}\boldsymbol{v}_1-\boldsymbol{C}\cdot(\gamma_1-\gamma_2)\boldsymbol{v}_1=0\end{cases}\\&\Leftrightarrow\begin{cases}\gamma_1\boldsymbol{C}^\top\boldsymbol{v}_2=(\gamma_1-\gamma_2)\boldsymbol{v}_1;\\(1-\gamma_2)\boldsymbol{C}\boldsymbol{v}_1=0.\end{cases}\end{aligned}$$

Since $\gamma_2\neq1$, the last equation implies $\boldsymbol{C}\boldsymbol{v}_1=0$. Left multiply the equality $\gamma_1\boldsymbol{C}^\top\boldsymbol{v}_2=(\gamma_1-\gamma_2)\boldsymbol{v}_1$ by $\boldsymbol{C}$ and use $\boldsymbol{C}\boldsymbol{v}_1=0$, and we obtain $\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{v}_2=0$, that is $\boldsymbol{C}^\top\boldsymbol{v}_2=0$. Substitute this back and we get $(\gamma_1-\gamma_2)\boldsymbol{v}_1=0$. But $\gamma_1\neq\gamma_2$, so $\boldsymbol{v}_1=0$. Thus, $\boldsymbol{M}(\gamma_1,\gamma_2)\boldsymbol{v}=(1-\gamma_2)\boldsymbol{v}$ implies $\boldsymbol{v}_1=0$ and $\boldsymbol{C}^\top\boldsymbol{v}_2=0$, and one verifies that the converse is also true. Therefore, $1-\gamma_2$ is an eigenvalue of $\boldsymbol{M}(\gamma_1,\gamma_2)$ with multiplicity n_2-r, that is the dimension of the nullspace of $\boldsymbol{C}^\top$.

Case 3.2: $\gamma_1=\gamma_2$**.** In this case, we consider the linear equations $\boldsymbol{M}(\gamma_1,\gamma_1)\boldsymbol{v}=(1-\gamma_1)\boldsymbol{v}$ in variable $\boldsymbol{v}:=[\boldsymbol{v}_1;\boldsymbol{v}_2]\in\mathbb{R}^{n_1+n_2}$. And we have

$$\begin{aligned}\boldsymbol{M}(\gamma_1,\gamma_1)\boldsymbol{v}=(1-\gamma_1)\boldsymbol{v}&\Leftrightarrow\begin{cases}\boldsymbol{C}^\top\boldsymbol{v}_2=0\\\gamma_1(1-\gamma_1)\boldsymbol{C}\boldsymbol{v}_1+\gamma_1^2\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{v}_2=0\end{cases}\\&\Leftrightarrow\boldsymbol{C}_2\boldsymbol{v}=0,\quad\text{where }\boldsymbol{C}_2:=\begin{bmatrix}\boldsymbol{C}&0\\0&\boldsymbol{C}^\top\end{bmatrix}.\end{aligned}$$

The nullspace of $\boldsymbol{C}_2$ is of dimension n_1+n_2-2r, so $1-\gamma_1$ is an eigenvalue of $\boldsymbol{M}(\gamma_1,\gamma_2)$ of multiplicity n_1+n_2-2r. The proof is now complete. □

Lemma 13 is important as it allows us to proceed for studying the spectral radius $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$. That is not to say analyzing $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$ is made into a trivial task. In fact, Table 2 already suggests an obstacle: If $n_1=r$ and $\gamma_1\neq1$, then $1-\gamma_1$ is not an eigenvalue of $\boldsymbol{M}(\gamma_1,\gamma_2)$ and should not be taken into account when studying $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$; similarly for $1-\gamma_2$. In other words, this suggests we have to further break our analysis into subcases that classify the relations between the rank r and block size n_1 (or n_2). Performing such analysis carefully is our main duty in Appendices B.2 and B.3.

B.2. Simplified Cases

In this section we make some basic calculations of the spectral radius in simplified cases. The purpose is to build some intuition and preliminary results for the sequel. The first easy case we consider is where $\gamma_1 = 1$ or $\gamma_2 = 1$. In this case, Lemma 13 reveals the whole spectrum of $\boldsymbol{M}(\gamma_1, \gamma_2)$, so we can minimize $\rho\big(\boldsymbol{M}(\gamma_1, \gamma_2)\big)$ with relative ease:

Proposition 1. *Recall $\boldsymbol{C} := \boldsymbol{A}_2^\top \boldsymbol{A}_1$ and $r := \operatorname{rank}(\boldsymbol{C})$. Under Assumption 1, $\rho\big(\boldsymbol{M}(\gamma_1, 1)\big)$ is minimized at $\gamma_1 = \frac{2}{2-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}$ if $r < n_1$, at $\gamma_1 = \frac{2}{2-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)}$ if $r = n_1 > 1$, and its minimum values are given as*

$$\min_{\gamma_1>0} \rho\big(\boldsymbol{M}(\gamma_1, 1)\big) = \begin{cases} \frac{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}{2-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)} & \text{if } r < n_1; \\ \frac{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)}{2-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)} & \text{if } r = n_1. \end{cases}$$

Remark 6. Similarly, we can solve $\min_{\gamma_2>0} \rho\big(\boldsymbol{M}(1, \gamma_2)\big)$.

Remark 7. If $r = n_1 = 1$, then $\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) = \lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)$, in which case the minimum of $\rho\big(\boldsymbol{M}(\gamma_1, 1)\big)$ is 0.

Proof of Proposition 1. Lemma 13 implies

$$\rho\big(\boldsymbol{M}(\gamma_1, 1)\big) = \begin{cases} \max\Big\{|1-\gamma_1|, \max_{i=1,\dots,r}\big\{|1-\gamma_1\big(1-\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\big)|\big\}\Big\} & \text{if } r < n_1; \\ \max_{i=1,\dots,r}\big\{|1-\gamma_1\big(1-\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\big)|\big\} & \text{if } r = n_1. \end{cases}$$

The proof finishes with a basic and standard argument; see, e.g., Lemma 8. □

Another easy case is where $\gamma_1 = \gamma_2$. In this case, we need to minimize $\rho\big(\boldsymbol{M}(\gamma_1, \gamma_2)\big)$ only in a single variable γ_1:

Proposition 2. *Define $\gamma_1^* := \frac{2}{1+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}$. Under Assumption 1, we have $\gamma_1^* - 1 < \rho\big(\boldsymbol{M}(1,1)\big)$ and*

$$\min_{\gamma_1>0} \rho\big(\boldsymbol{M}(\gamma_1, \gamma_1)\big) = \rho\big(\boldsymbol{M}(\gamma_1^*, \gamma_1^*)\big) = \gamma_1^* - 1. \tag{18}$$

Remark 8. While in Proposition 1 the minima can vary with the rank r, the minimum of (18) is attained at γ_1^* for any r.

Proof of Proposition 2. Under Assumption 1, γ_1^* is real-valued, and with some calculation, one further verifies $0 < \gamma_1^* - 1 < \lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) = \rho\big(\boldsymbol{M}(1,1)\big)$; the last equality is due to Example 1. We proceed in two steps. In Step 1 we prove $\rho\big(\boldsymbol{M}(\gamma_1^*, \gamma_1^*)\big) = \gamma_1^* - 1$. In Step 2 we prove $\gamma_1^* - 1 = \min_{\gamma_1>0} \rho\big(\boldsymbol{M}(\gamma_1, \gamma_1)\big)$.

Step 1. First observe that for every $i = 1, \dots, r$ we have

$$\begin{aligned} &\Big(2 - 2\gamma_1^* + (\gamma_1^*)^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\Big)^2 - 4(\gamma_1^* - 1)^2 \le 0 \\ \Leftrightarrow\ &(\gamma_1^*)^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top) \cdot \Big(4 - 4\gamma_1^* + (\gamma_1^*)^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\Big) \le 0 \\ \Leftrightarrow\ &4 - 4\gamma_1^* + (\gamma_1^*)^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top) \le 0. \end{aligned}$$

But as one can verify, the definition of γ_1^* implies γ_1^* is the smaller root of the equation $4 - 4\xi + \xi^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) = 0$ in variable ξ, that is $4 - 4\gamma_1^* + (\gamma_1^*)^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) = 0$. As a consequence, for every $i = 1, \dots, r$, the roots of the quadratic equation

$$z^2 - \Big(2 - 2\gamma_1^* + (\gamma_1^*)^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\Big)z + (\gamma_1^* - 1)^2 = 0$$

in variable z have magnitudes equal to $\gamma_1^* - 1$. Invoking Lemma 13, we see that these roots are precisely $2r$ eigenvalues $\boldsymbol{M}(\gamma_1^*, \gamma_1^*)$, while Lemma 13 further suggests that $\boldsymbol{M}(\gamma_1^*, \gamma_1^*)$ might have eigenvalue $\gamma_1^* - 1$ with multiplicity $n_1 + n_2 - 2r$ (if $n_1 + n_2 - 2r \neq 0$). We can now conclude *Step 1* with $\rho\big(\boldsymbol{M}(\gamma_1^*, \gamma_1^*)\big) = \gamma_1^* - 1$.

Step 2. To prove $\gamma_1^* - 1 = \min_{\gamma_1>0} \rho\big(\boldsymbol{M}(\gamma_1, \gamma_1)\big)$, we consider two cases.

Case 2.1: $\gamma_1 > \gamma_1^*$. In this case, $\gamma_1 > \gamma_1^* > 1$. Note that for every $i = 1, \dots, r$ the equation (16) must has a root whose magnitude is larger than or equal to $\sqrt{(1-\gamma_1)(\gamma_2 - 1)} = \gamma_1 - 1$. Therefore $\rho\big(\boldsymbol{M}(\gamma_1, \gamma_1)\big) \ge \gamma_1 - 1 > \gamma_1^* - 1$.

Case 2.2: $0 < \gamma_1 \leq \gamma_1^*$**.** The choice of γ_1^* indicates $4-4\gamma_1+\gamma_1^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) \geq 0$. We then prove $2-2\gamma_1+\gamma_1^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) \geq 0$: This is true if $\gamma_1 \in (0,1]$; otherwise, if $\gamma_1 > 1$, this is implied by $4-4\gamma_1+\gamma_1^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) > 0$. As a result, the quadratic equation $z^2 - \big(2-2\gamma_1+\gamma_1^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)\big)z+(\gamma_1-1)^2 = 0$ has two different non-negative real roots and they are eigenvalues of $\boldsymbol{M}(\gamma_1,\gamma_1)$. The larger root, denoted by $z(\gamma_1)$, is given as

$$\begin{aligned} z(\gamma_1) &= \frac{2-2\gamma_1+\gamma_1^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)+\sqrt{\big(2-2\gamma_1+\gamma_1^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)\big)^2-4(\gamma_1-1)^2}}{2} \\ &= \frac{\Big(\gamma_1\sqrt{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}+\sqrt{4-4\gamma_1+\gamma_1^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}\Big)^2}{4}. \end{aligned}$$

We need to prove $z(\gamma_1) \geq \gamma_1^* - 1$ for every $\gamma_1 \in (0,\gamma_1^*]$. Since $z(\gamma_1^*) = \gamma_1^* - 1$, it suffices to show that the function

$$f(\xi) = \xi\sqrt{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}+\sqrt{4-4\xi+\xi^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}$$

is non-increasing in $[0,\gamma_1^*]$, i.e., its derivative is non-positive. Noting that, with $\xi \in [0,\gamma_1^*]$ we have $2-\xi\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) \geq 2-\gamma_1^*\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) \geq 0$. We can then verify

$$\begin{aligned} f'(\xi) \leq 0 &\Leftrightarrow \sqrt{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}+\frac{\xi\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)-2}{\sqrt{4-4\xi+\xi^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}} \leq 0 \\ &\Leftrightarrow \lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)\cdot\Big(4-4\xi+\xi^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)\Big) \leq \Big(2-\xi\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)\Big)^2 \\ &\Leftrightarrow \lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) \leq 1. \end{aligned}$$

We have thus completed the proof. □

From Propositions 1 and 2, we see that, under Assumption 1, stepsizes better than $\gamma_1 = \gamma_2 = 1$ do exist, suggesting that it is possible for BGD to converge faster than BEM.

Note that the minima in Propositions 1 and 2 are both attained at some stepsize larger than 1, behind which the intuition is as follows. Suppose Assumption 1 holds, then the maximum eigenvalue of $\boldsymbol{C}\boldsymbol{C}^\top$ is smaller than or equal to 1. If furthermore $\gamma_1 = 1$, then $\boldsymbol{M}(\gamma_1,\gamma_2)$ can be written as $\boldsymbol{M}(1,\gamma_2) = \begin{bmatrix} 0 & -\boldsymbol{C}^\top \\ 0 & (1-\gamma_2)\boldsymbol{I}_2+\gamma_2\boldsymbol{C}\boldsymbol{C}^\top \end{bmatrix}$. If γ_2 were restricted to lie in $(0,1]$, then there would be no doubt that the convex combination $(1-\gamma_2)\boldsymbol{I}_2+\gamma_2\boldsymbol{C}\boldsymbol{C}^\top$ would be minimized at $\gamma_2 = 1$ under Assumption 1. However, inspect that setting γ_2 to be appropriately larger than 1 would actually further reduce the magnitudes of $(1-\gamma_2)\boldsymbol{I}_2+\gamma_2\boldsymbol{C}\boldsymbol{C}^\top$ and therefore of its eigenvalues. Optimal stepsizes that we show in Appendix B.3 are in fact all larger than 1.

B.3. General Cases

The limitation of Propositions 1 and 2 is that, there, we search stepsizes on two rays, namely $\{(\gamma_1,1):\gamma_1>0\}$ and $\{(\gamma_1,\gamma_1):\gamma_1>0\}$, while a better choice might lie on the *quadrant* $\{(\gamma_1,\gamma_2):\gamma_1>0,\gamma_2>0\}$ yet on neither of the two rays. To address this point, we divide the quadrant into four regions, $S_{00},S_{01},S_{10},S_{11}$ as shown in Figure 2, and we search for stepsizes that leading to smaller spectral radii over each region. Combining yields a solution to (3).

To proceed, we develop a technical lemma, shown below, which will be used as a common sub-routine to prove our main results in Appendices B.3.1 and B.3.2.

Lemma 14. *Fix $\phi \in (0,1)$. Consider the following quadratic equation in variable z and its discriminant:*

$$\begin{aligned} &z^2-(2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi)z+(1-\gamma_1)(1-\gamma_2)=0, \\ &\Delta(\gamma_1,\gamma_2) := (2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi)^2 \\ &\qquad\qquad -4(1-\gamma_1)(1-\gamma_2). \end{aligned} \tag{19}$$

Assume $\Delta(\gamma_1,\gamma_2) \geq 0$. Then

$$\begin{cases} 2\sqrt{\Delta(\gamma_1,\gamma_2)}\cdot|\gamma_1\phi-1| \leq \left|\frac{\partial\Delta(\gamma_1,\gamma_2)}{\partial\gamma_2}\right| & \text{if } \gamma_1 \geq 1; \\ 2\sqrt{\Delta(\gamma_1,\gamma_2)}\cdot|\gamma_1\phi-1| \geq \left|\frac{\partial\Delta(\gamma_1,\gamma_2)}{\partial\gamma_2}\right| & \text{if } \gamma_1 \in (0,1). \end{cases}$$

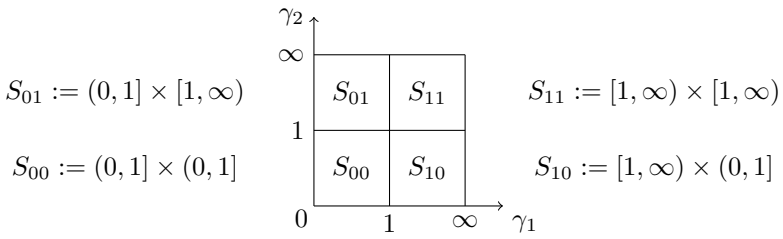


Figure 2: We divide the quadrant $\{(\gamma_1,\gamma_2) : \gamma_1 > 0, \gamma_2 > 0\}$ of all possible stepsizes into 4 regions, S_{00}, S_{01}, S_{10}, and S_{11}. We minimize the spectral radius $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$ over each region separately, which will give a solution to (3).

Remark 9. In words, Lemma 14 analyzes a single quadratic equation (19) which is of the form (16), highlighting some properties useful for analyzing the eigenvalues as roots of (16) in the sequel.

Proof of Lemma 14. We only write down the proof for the case $\gamma_1 \geq 1$; the case for $\gamma_1 \in (0,1)$ is similar. First note that

$$2\sqrt{\Delta(\gamma_1,\gamma_2)} \cdot |\gamma_1\phi - 1| \leq \big|\frac{\partial \Delta(\gamma_1,\gamma_2)}{\partial \gamma_2}\big| \Leftrightarrow 4\Delta(\gamma_1,\gamma_2) \cdot (\gamma_1\phi - 1)^2 \leq \Big(\frac{\partial \Delta(\gamma_1,\gamma_2)}{\partial \gamma_2}\Big)^2.$$

Canceling a common additive term yields the following equivalent inequality:

$$-(\gamma_1-1)(\gamma_2-1)(\gamma_1\phi-1)^2 \leq (\gamma_1-1)^2 - (2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi)(\gamma_1\phi-1)(\gamma_1-1).$$

If $\gamma_1 = 1$, we are done. If $\gamma_1 > 1$, we can divide both sides of the above by $\gamma_1 - 1$ to get an equivalent inequality:

$$\begin{aligned}
&-(\gamma_2-1)(\gamma_1\phi-1)^2 \leq (\gamma_1-1) - (2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi)(\gamma_1\phi-1) \\
\Leftrightarrow\ &(\gamma_1\phi-1)\cdot\big((2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi) - (\gamma_2-1)(\gamma_1\phi-1)\big) \leq (\gamma_1-1) \\
\Leftrightarrow\ &(\gamma_1\phi-1)\cdot(1-\gamma_1+\gamma_1\phi) \leq (\gamma_1-1) \\
\Leftrightarrow\ &\gamma_1^2\phi(\phi-1) \leq 0.
\end{aligned}$$

But $\phi \in (0,1)$, so $\gamma_1^2\phi(\phi-1) \leq 0$, and we finished the proof. □

B.3.1. Optimal Stepsizes on $S_{00} \cup S_{01} \cup S_{10}$

Let us first acquire a better understanding of how the roots of a single quadratic equation in (19) behave:

Lemma 15. *Fix $\phi \in (0,1)$. Consider the quadratic equation in (19) and let $z_1(\gamma_1,\gamma_2)$, $z_2(\gamma_1,\gamma_2)$ be its two roots. Define*

$$f(\gamma_1,\gamma_2) := \max\big\{|z_1(\gamma_1,\gamma_2)|, |z_2(\gamma_1,\gamma_2)|\big\}.$$

With S_{00}, S_{01}, S_{10} defined in Figure 2, the following hold:

- *Part 1: S_{00}. With $\gamma_1 \in (0,1]$ fixed, $\gamma_2 \mapsto f(\gamma_1,\gamma_2)$ is non-increasing in $(0,1]$. With $\gamma_2 \in (0,1]$ fixed, $\gamma_1 \mapsto f(\gamma_1,\gamma_2)$ is non-increasing in $(0,1]$. Therefore*

$$\min_{[\gamma_1;\gamma_2]\in S_{00}} f(\gamma_1,\gamma_2) = f(1,1).$$

- *Part 2: $S_{01} \cup S_{10}$. With $\gamma_2 \in [1,\infty)$ fixed, $\gamma_1 \mapsto f(\gamma_1,\gamma_2)$ is non-increasing in $(0,1]$. With $\gamma_1 \in [1,\infty)$ fixed, $\gamma_2 \mapsto f(\gamma_1,\gamma_2)$ is non-increasing in $(0,1]$. Hence we have*

$$\begin{aligned}
\min_{[\gamma_1;\gamma_2]\in S_{01}} f(\gamma_1,\gamma_2) &= \min_{\gamma_2\geq 1} f(1,\gamma_2), \\
\min_{[\gamma_1;\gamma_2]\in S_{10}} f(\gamma_1,\gamma_2) &= \min_{\gamma_1\geq 1} f(\gamma_1,1).
\end{aligned}$$

Proof of Lemma 15. Recall $\Delta(\gamma_1,\gamma_2) := (2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi)^2 - 4(1-\gamma_1)(1-\gamma_2)$ (Lemma 14). We present the proofs below for *Part 1* and *Part 2*, one after another.

Part 1: S_{00}**.** Assume $\gamma_1, \gamma_2 \in (0, 1]$, and we derive the monotonicity of f. In this case we have $1 - \gamma_1 \geq 0$ and $1 - \gamma_2 \geq 0$, and therefore we can write

$$\Delta(\gamma_1, \gamma_2) = \left(\left(\sqrt{1-\gamma_1} - \sqrt{1-\gamma_2}\right)^2 + \gamma_1\gamma_2\phi\right)\left(\left(\sqrt{1-\gamma_1} + \sqrt{1-\gamma_2}\right)^2 + \gamma_1\gamma_2\phi\right),$$

which implies $\Delta(\gamma_1, \gamma_2) > 0$. Furthermore, since $2 - \gamma_1 - \gamma_2 + \gamma_1\gamma_2\phi \geq 0$, we have

$$\frac{\partial f(\gamma_1, \gamma_2)}{\partial \gamma_1} = \frac{-1 + \gamma_2\phi}{2} + \frac{1}{4\sqrt{\Delta(\gamma_1, \gamma_2)}} \cdot \frac{\partial \Delta(\gamma_1, \gamma_2)}{\partial \gamma_1}.$$

Since $\gamma_2 \in (0, 1)$, applying Lemma 14 with γ_1 and γ_2 swapped yields $2\sqrt{\Delta(\gamma_1, \gamma_2)} \cdot |\gamma_1\phi - 1| \geq \left|\frac{\partial \Delta(\gamma_1, \gamma_2)}{\partial \gamma_1}\right|$. Since $-1 + \gamma_2\phi \leq 0$, this further implies $\frac{\partial f(\gamma_1, \gamma_2)}{\partial \gamma_1} \leq 0$. Similarly, using Lemma 14 we can show $\frac{\partial f(\gamma_1, \gamma_2)}{\partial \gamma_2} \leq 0$. This proves *Part 1*.

Part 2: $S_{01} \cup S_{10}$**.** Since $(1 - \gamma_1)(1 - \gamma_2) \leq 0$, the two roots of (19) are real-valued. It suffices to fix $\gamma_1 \in [1, \infty)$ and prove $\gamma_2 \mapsto f(\gamma_1, \gamma_2)$ is non-increasing in $(0, 1]$. To do so, we first calculate $\frac{\partial \Delta(\gamma_1, \gamma_2)}{\partial \gamma_2}$ and verify $\frac{\partial \Delta(\gamma_1, \gamma_2)}{\partial \gamma_2} \leq 0$ for every $\gamma_2 \in (0, 1]$:

$$\begin{aligned}
\frac{\partial \Delta(\gamma_1, \gamma_2)}{\partial \gamma_2} \leq 0 &\Leftrightarrow 2(\gamma_1\phi - 1)^2\gamma_2 - 2(\gamma_1\phi + 1 - 2\phi)\gamma_1 \leq 0 \\
&\Leftarrow (\gamma_1\phi - 1)^2 - (\gamma_1\phi + 1 - 2\phi)\gamma_1 \leq 0 \\
&\Leftrightarrow \gamma_1^2\phi(\phi - 1) + 1 - \gamma_1 \leq 0.
\end{aligned}$$

This indeed holds as $\phi \in (0, 1)$ and $\gamma_1 \geq 1$. Then, by definition we can write

$$\frac{\partial f(\gamma_1, \gamma_2)}{\partial \gamma_2} = \begin{cases} \frac{-1+\gamma_1\phi}{2} + \frac{1}{4\sqrt{\Delta(\gamma_1, \gamma_2)}} \cdot \frac{\partial \Delta(\gamma_1, \gamma_2)}{\partial \gamma_2} & \text{if } 2 - \gamma_1 - \gamma_2 + \gamma_1\gamma_2\phi > 0; \\ \frac{1-\gamma_1\phi}{2} + \frac{1}{4\sqrt{\Delta(\gamma_1, \gamma_2)}} \cdot \frac{\partial \Delta(\gamma_1, \gamma_2)}{\partial \gamma_2} & \text{if } 2 - \gamma_1 - \gamma_2 + \gamma_1\gamma_2\phi < 0; \\ \sqrt{(\gamma_1 - 1)(1 - \gamma_2)} & \text{if } 2 - \gamma_1 - \gamma_2 + \gamma_1\gamma_2\phi = 0. \end{cases}$$

Here, $\gamma_2 \mapsto (\gamma_1 - 1)(1 - \gamma_2)$ is clearly a decreasing function. For the other two cases, we can also prove $\frac{\partial f(\gamma_1, \gamma_2)}{\partial \gamma_2} \leq 0$ by using Lemma 14 and the fact $\frac{\partial \Delta(\gamma_1, \gamma_2)}{\partial \gamma_2} \leq 0$. □

Having studied a single quadratic equation in Lemma 15, we can now analyze r such equations (16). Specifically, armed with Lemmas 14 and 15, we reach the following result:

Theorem 3. *Under Assumption 1, the following hold.*

- *(Part 1:* S_{00}*) On* S_{00}*, we have*

$$\min_{[\gamma_1;\gamma_2] \in S_{00}} \rho\big(\boldsymbol{M}(\gamma_1, \gamma_2)\big) = \rho\big(\boldsymbol{M}(1, 1)\big).$$

- *(Part 2:* $S_{01} \cup S_{10}$*) We have*

$$\begin{aligned}
\min_{[\gamma_1;\gamma_2] \in S_{01}} \rho\big(\boldsymbol{M}(\gamma_1, \gamma_2)\big) &= \min_{\gamma_2 \geq 1} \rho\big(\boldsymbol{M}(1, \gamma_2)\big), \\
\min_{[\gamma_1;\gamma_2] \in S_{10}} \rho\big(\boldsymbol{M}(\gamma_1, \gamma_2)\big) &= \min_{\gamma_1 \geq 1} \rho\big(\boldsymbol{M}(\gamma_1, 1)\big).
\end{aligned}$$

Remark 10. The situation on S_{00} is now clear (*Part 1*): The minimum spectral radius is attained at $(1, 1)$. The situation on $S_{01} \cup S_{10}$ is clearly only partially (*Part 2*), and it seems that one still needs to minimize over a single stepsize γ_2 (*resp.* γ_1) to find the minimum spectral radius on S_{01} (*resp.* S_{10}), with the other stepsize set to 1. This is in fact an easier task and has been addressed already in Proposition 1.

Proof of Theorem 3. The proof of Theorem 3 relies on two technical lemmas, Lemmas 14 and 15.

Let $z_{i1}(\gamma_1, \gamma_2)$ and $z_{i2}(\gamma_1, \gamma_2)$ be the two roots of (16). Define

$$f_i(\gamma_1, \gamma_2) := \max\big\{|z_{i1}(\gamma_1, \gamma_2)|, |z_{i2}(\gamma_1, \gamma_2)|\big\}. \tag{20}$$

If $r < \min\{n_1, n_2\}$, then, by Lemma 13, $1-\gamma_1$ and $1-\gamma_2$ are eigenvalues of $\boldsymbol{M}(\gamma_1,\gamma_2)$, and Lemmas 9 and 15 further imply

$$\begin{aligned}\min_{[\gamma_1;\gamma_2]\in S_{00}} \rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big) &= \min_{[\gamma_1;\gamma_2]\in S_{00}} \left(\max\left\{1-\gamma_1, 1-\gamma_2, \max_{i,\dots,r} f_i(\gamma_1,\gamma_2)\right\}\right) \\ &= \rho\big(\boldsymbol{M}(1,1)\big).\end{aligned}$$

The case where $r = n_1$ or $r = n_2$ is similar, and this finishes the proof of *Part 1*. *Part 2* follows similarly from Lemma 9 and *Part 2* of Lemma 15. □

B.3.2. Optimal Stepsizes on S_{11}

Solving (3) on S_{11} is harder than on $S_{00} \cup S_{01} \cup S_{10}$, as on S_{11} it is more challenging to quantify the change of the roots (eigenvalues) defined in (16) with respect to stepsize γ_1 or γ_2. We tackle this challenge by further dividing our analysis into two separate cases: $\boldsymbol{C}$ has full rank and $\boldsymbol{C}$ is rank-deficient.

The proofs for the two cases are much more sophisticated. Hence we only state the theorems below. The proofs are presented separately in Appendix C.

The Full Rank Case. If $\boldsymbol{C}$ is full rank (i.e., $r = \min\{n_1, n_2\}$), we have:

Theorem 4. *Suppose $r = \min\{n_1, n_2\}$ and Assumption 1 holds. Let S_{11} be defined in Figure 2. Then the minimum value $\min_{[\gamma_1;\gamma_2]\in S_{11}} \rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$ is equal to*

$$\frac{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)} - \sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)} + \sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}. \tag{21}$$

Remark 11. If $r = 1$, then we have $\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) = \lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)$ and Theorem 4 suggests the minimum spectral radius is 0, coinciding with Proposition 1 and Remark 7.

The Rank-Deficient Case. It remains to address the case where $\boldsymbol{C}$ is rank-deficient (i.e., $r < \min\{n_1, n_2\}$). We do this in Theorem 5:

Theorem 5. *Suppose $r < \min\{n_1, n_2\}$ and Assumption 1 holds. With S_{11} defined in Figure 2 and γ_1^* defined in Proposition 2, we have $\rho\big(\boldsymbol{M}(\gamma_1^*,\gamma_1^*)\big) = \gamma_1^* - 1$ and*

$$\begin{aligned}\min_{[\gamma_1;\gamma_2]\in S_{11}} \rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big) &= \gamma_1^* - 1 \\ &= \frac{1-\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}{1+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}.\end{aligned} \tag{22}$$

There is a basic connection between (21) and (22): They would be identical if $\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)$ were equal to 0. Despite this direct connection, their proofs are very different.

C. Proofs of Theorems 4 and 5

C.1. Proof of Theorem 4

For the proof of Theorem 4, we need the following lemma.

Lemma 16. *Assume $s > 1$ and $\zeta_1 \geq \cdots \geq \zeta_s > 0$. For $i = 1, \dots, s$, let $z_{i1}(\alpha,\beta)$, $z_{i2}(\alpha,\beta)$ be the two roots of the following quadratic equation in variable z:*

$$z^2 - (\beta + 1 - \alpha\zeta_i)z + \beta = 0. \tag{23}$$

Define $g_i(\alpha,\beta) := \max\{|z_{i1}(\alpha,\beta)|, |z_{i2}(\alpha,\beta)|\}$. Then

$$\min_{\alpha>0,\beta\geq 0} \Big(\max_{i=1,\dots,s} g_i(\alpha,\beta)\Big) = \frac{\sqrt{\zeta_1}-\sqrt{\zeta_s}}{\sqrt{\zeta_1}+\sqrt{\zeta_s}},$$

where the minimum is attained at

$$\alpha = \Big(\frac{2}{\sqrt{\zeta_1}+\sqrt{\zeta_s}}\Big)^2, \quad \beta = \Big(\frac{\sqrt{\zeta_1}-\sqrt{\zeta_s}}{\sqrt{\zeta_1}+\sqrt{\zeta_s}}\Big)^2.$$

Proof of Lemma 16. By the definition of (23) we have

$$|z_{i1}(\alpha,\beta)| \cdot |z_{i2}(\alpha,\beta)| = \beta.$$

Without loss of generality, we assume $|z_{i1}(\alpha,\beta)| \geq |z_{i2}(\alpha,\beta)|$ If $z_{i1}(\alpha,\beta)$ and $z_{i1}(\alpha,\beta)$ are two complex roots, then their magnitudes are equal to $\sqrt{\beta}$. If they are real roots, then we must have $|z_{i1}(\alpha,\beta)| \geq \sqrt{\beta}$, where the equality is attained if and only if the two real roots are equal. Therefore $g_i(\alpha,\beta) \geq \sqrt{\beta}$. As a consequence, for any $\beta \geq 0$ the optimization problem

$$\min_{\alpha>0}\Big(\max_{i=1,\ldots,s} g_i(\alpha,\beta)\Big) \tag{24}$$

is lower bounded by $\sqrt{\beta}$. Clearly, the lower bound $\sqrt{\beta}$ of (24) is attained at α if and only if α satisfies the following for all $i = 1,\ldots,s$:

$$\begin{aligned} &(\beta+1-\alpha\zeta_i)^2 - 4\beta \leq 0 \\ \Leftrightarrow\ &\big(\beta+1-\alpha\zeta_i+2\sqrt{\beta}\big)\big(\beta+1-\alpha\zeta_i-2\sqrt{\beta}\big) \leq 0 \\ \Leftrightarrow\ &\big((\sqrt{\beta}+1)^2-\alpha\zeta_i\big)\big((\sqrt{\beta}-1)^2-\alpha\zeta_i\big) \leq 0 \\ \Leftrightarrow\ &\big|1-\sqrt{\alpha\zeta_i}\big| \leq \sqrt{\beta} \leq 1+\sqrt{\alpha\zeta_i}. \end{aligned} \tag{25}$$

Next, for β to be as small as possible while satisfying (25), we need to choose α to be the global minimizer of

$$\min_{\alpha>0}\Big\{\max_{i=1,\ldots,s}\big|1-\sqrt{\alpha\zeta_i}\big|\Big\} = \min_{\alpha>0}\Big\{\max\big\{\big|1-\sqrt{\alpha\zeta_1}\big|,\big|1-\sqrt{\alpha\zeta_s}\big|\big\}\Big\}.$$

A simple geometric argument shows that the minimum of the above problem is attained at $\alpha = \frac{4}{(\sqrt{\zeta_1}+\sqrt{\zeta_s})^2}$, with the associated minimum value being $\frac{\sqrt{\zeta_1}-\sqrt{\zeta_s}}{\sqrt{\zeta_1}+\sqrt{\zeta_s}}$. To summarize, with $\alpha^* = \frac{4}{(\sqrt{\zeta_1}+\sqrt{\zeta_s})^2}$ and $\beta^* = \frac{(\sqrt{\zeta_1}-\sqrt{\zeta_s})^2}{(\sqrt{\zeta_1}+\sqrt{\zeta_s})^2}$, we have $\max_{i=1,\ldots,s} g_i(\alpha^*,\beta^*)$ equal to $\sqrt{\beta^*}$.

To prove optimality, we need to show that for any α,β that satisfies $\sqrt{\beta} < \big|1-\sqrt{\alpha\zeta_i}\big|$ for some i, the corresponding objective is larger than $\sqrt{\beta^*}$. To do so, it suffices to consider the following cases:

- If $\alpha > \alpha^*$, then $|1-\sqrt{\alpha\zeta_s}| < |1-\sqrt{\alpha\zeta_1}|$. Assume $\sqrt{\beta} < |1-\sqrt{\alpha\zeta_1}|$. The equation $z^2-(\beta+1-\alpha\zeta_1)z+\beta=0$ has two real roots, and the larger one is

  $$\frac{|\beta+1-\alpha\zeta_1|+\sqrt{(\beta+1-\alpha\zeta_1)^2-4\beta}}{2}.$$

  The derivative of this root with respect to β is

  $$\pm\frac{1}{2}+\frac{\beta-\alpha\zeta_1-1}{2\sqrt{(\beta+1-\alpha\zeta_1)^2-4\beta}},$$

  which is negative regardless of the sign on $1/2$ (use the assumption $\sqrt{\beta} < |1-\sqrt{\alpha\zeta_i}|$ to verify). Hence the root is minimized when $\sqrt{\beta} = |1-\sqrt{\alpha\zeta_1}|$, at which point this root can be written as

  $$\frac{(1-\sqrt{\alpha\zeta_1})^2+1-\alpha\zeta_1}{2} = |1-\sqrt{\alpha\zeta_1}|.$$

  Since $\alpha > \alpha^*$, one verifies $|1-\sqrt{\alpha\zeta_1}|$ is larger than $|1-\sqrt{\alpha^*\zeta_1}| = \sqrt{\beta^*}$.

- The case $\alpha < \alpha^*$ can be proved similarly. Specifically, in this case we have $|1-\sqrt{\alpha\zeta_1}| < |1-\sqrt{\alpha\zeta_s}|$, so we can assume $\sqrt{\beta} < |1-\sqrt{\alpha\zeta_s}|$ and proceed with a similar argument.

□

We are then ready to prove Theorem 4.

Proof of Theorem 4. We can easily verify from Lemma 13 that

$$\rho\Big(\boldsymbol{M}\Big(\frac{1}{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)},1\Big)\Big)=0 \quad \text{if } r=n_1=1;$$
$$\rho\Big(\boldsymbol{M}\Big(1,\frac{1}{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}\Big)\Big)=0 \quad \text{if } r=n_2=1.$$

Next we assume $r>1$. To proceed, we consider the function $\tau: S_{11}\to\mathbb{R}^2$ defined as

$$\tau(\gamma_1,\gamma_2):=[\gamma_1\gamma_2;(\gamma_1-1)(\gamma_2-1)]. \tag{26}$$

Then we can write the image $\tau(S_{11})$ as

$$\tau(S_{11}):=\big\{[\gamma_1\gamma_2;(\gamma_1-1)(\gamma_2-1)]\in\mathbb{R}^2:[\gamma_1;\gamma_2]\in S_{11}\big\}.$$

Note that we have $\alpha>0$ and $\beta\geq 0$ for every $[\alpha,\beta]\in\tau(S_{11})$.

On the other hand, observe that (16) can be written as

$$z^2-\Big((\gamma_1-1)(\gamma_2-1)+1-\gamma_1\gamma_2\big(1-\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\big)\Big)z+(\gamma_1-1)(\gamma_2-1)=0. \tag{27}$$

By inspecting (23) of Lemma 16 and (27), we find we can invoke Lemma 16 with $s=r$ and $\zeta_i=1-\lambda_{r+1-i}(\boldsymbol{C}\boldsymbol{C}^\top)$ and obtain

$$\begin{aligned}\min_{[\gamma_1;\gamma_2]\in S_{11}}\Big(\max_{i=1,\ldots,r} f_i(\gamma_1,\gamma_2)\Big)&=\min_{[\alpha,\beta]\in\tau(S_{11})}\Big(\max_{i=1,\ldots,r} g_i(\alpha,\beta)\Big)\\&\geq\min_{\alpha>0,\beta\geq 0}\Big(\max_{i=1,\ldots,r} g_i(\alpha,\beta)\Big)\\&=\max_{i=1,\ldots,r} g_i(\alpha^*,\beta^*)\\&=\frac{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)}-\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)}+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}},\end{aligned} \tag{28}$$

where the last two steps follow from Lemma 16 with α^* and β^* defined as

$$\alpha^*=\Big(\frac{2}{\sqrt{\zeta_1}+\sqrt{\zeta_r}}\Big)^2=\left(\frac{2}{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)}+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}\right)^2,$$
$$\beta^*=\Big(\frac{\sqrt{\zeta_1}-\sqrt{\zeta_r}}{\sqrt{\zeta_1}+\sqrt{\zeta_r}}\Big)^2=\left(\frac{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)}-\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}{\sqrt{1-\lambda_r(\boldsymbol{C}\boldsymbol{C}^\top)}+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}\right)^2.$$

We will show $[\alpha^*;\beta^*]\in\tau(S_{11})$, as this will prove the inequality in (28) is in fact an equality. Indeed, we can always solve the equations

$$\gamma_1\gamma_2=\alpha^*,\quad (\gamma_1-1)(\gamma_2-1)=\beta^*$$

for γ_1 and γ_2 and obtain two solutions $[\gamma_1^*;\gamma_2^*]\in S_{11}$ and $[\gamma_2^*;\gamma_1^*]\in S_{11}$ (as the reader could verify), where γ_1^* and γ_2^* are defined as

$$\gamma_1^*=\left(\frac{\sqrt{(1+\sqrt{\zeta_r})(1+\sqrt{\zeta_1})}+\sqrt{(1-\sqrt{\zeta_r})(1-\sqrt{\zeta_1})}}{\sqrt{\zeta_1}+\sqrt{\zeta_r}}\right)^2,$$
$$\gamma_2^*=\left(\frac{\sqrt{(1+\sqrt{\zeta_r})(1+\sqrt{\zeta_1})}-\sqrt{(1-\sqrt{\zeta_r})(1-\sqrt{\zeta_1})}}{\sqrt{\zeta_1}+\sqrt{\zeta_r}}\right)^2.$$

Of course $(\gamma_1^* - 1)(\gamma_2^* - 1) = \beta^*$ by construction, so we have proved

$$\min_{[\gamma_1;\gamma_2]\in S_{11}} \Big(\max_{i=1,\dots,r} f_i(\gamma_1,\gamma_2)\Big) = \max_{i=1,\dots,r} f_i(\gamma_1^*,\gamma_2^*) = \max_{i=1,\dots,r} f_i(\gamma_2^*,\gamma_1^*) = \sqrt{\beta^*}.$$

We can now finish the proof by inspecting Lemma 13 and the values of γ_1^* and γ_2^*. Specifically, we have the following three cases.

- If $r = n_1 = n_2$, then neither $1-\gamma_1^*$ nor $1-\gamma_2^*$ is an eigenvalue of $\boldsymbol{M}(\gamma_1^*,\gamma_2)$, therefore both $[\gamma_1^*;\gamma_2^*]$ and $[\gamma_2^*;\gamma_1^*]$ minimize $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$ on S_{11}.

- If $r = n_1 < n_2$, then $\boldsymbol{M}(\gamma_1^*,\gamma_2^*)$ has one extra eigenvalue $1-\gamma_2^*$. Since $\gamma_1^* \ge \gamma_2^* \ge 1$ and $\beta^* = (\gamma_1^*-1)(\gamma_2^*-1) \ge (\gamma_2^*-1)^2$, we have

  $$\min_{[\gamma_1;\gamma_2]\in S_{11}} \Big(\max_{i=1,\dots,r} f_i(\gamma_1,\gamma_2)\Big) = \max_{i=1,\dots,r} f_i(\gamma_1^*,\gamma_2^*) = \sqrt{\beta^*} \ge \gamma_2^* - 1,$$

  and therefore $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$ is minimized at $[\gamma_1^*;\gamma_2^*]$ on S_{11}.

- If $r = n_2 < n_1$, similarly, $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$ is minimized at $[\gamma_2^*;\gamma_1^*]$ on S_{11}.

□

C.2. Proof of Theorem 5

Here we prove Theorem 5. This is achieved by extending Lemma 14 into an analysis of the monotonicity of the determinant $\Delta(\gamma_1,\gamma_2)$ in (19). We begin with stating and proving Lemmas 17 and 18.

Lemma 17. *Let $\phi \in (0,1)$ and $\gamma_1 \in [1,2)$ be fixed with $\gamma_1\phi \ne 1$. Consider*

$$\Delta(\gamma_2) = (2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi)^2 - 4(\gamma_1-1)(\gamma_2-1).$$

Let ξ_1 and ξ_2 be the two roots of $\Delta(\gamma_2)=0$ with $|\xi_1| \le |\xi_2|$. Then $1 \le \xi_1$. Moreover, we have:

- *$\Delta'(\gamma_2) \le 0$ for every $\gamma_2 \in [1,\gamma_1]$.*
- *If $\gamma_1 \in [1, \frac{2}{1+\sqrt{1-\phi}}]$, then $\Delta(\gamma_2) \ge 0$ for every $\gamma_2 \in [1,\gamma_1]$.*
- *If $\gamma_1 \in [\frac{2}{1+\sqrt{1-\phi}}, 2)$, then $\xi_1 \le \gamma_1$ and*

$$\begin{aligned} &\Delta(\gamma_2) \ge 0, \quad \forall \gamma_2 \in [1,\xi_1], \\ &\Delta(\gamma_2) \le 0, \quad \forall \gamma_2 \in [\xi_1,\gamma_1]. \end{aligned}$$

Proof. Since $\gamma_1\phi \ne 1$, we know $\Delta(\gamma_2)$ is a quadratic function in γ_2, and we can simplify its expression and calculate its derivative as follows:

$$\begin{aligned} \Delta(\gamma_2) &= (\gamma_1\phi-1)^2\gamma_2^2 + 2(2-\gamma_1)(\gamma_1\phi-1)\gamma_2 - 4(\gamma_1-1)\gamma_2 + (2-\gamma_1)^2 + 4(\gamma_1-1) \\ &= (\gamma_1\phi-1)^2\gamma_2^2 - 2(\gamma_1\phi+1-2\phi)\gamma_1\gamma_2 + \gamma_1^2, \\ \Delta'(\gamma_2) &= 2(2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi)(\gamma_1\phi-1) - 4(\gamma_1-1) \\ &= 2(\gamma_1\phi-1)^2\gamma_2 - 2(\gamma_1\phi+1-2\phi)\gamma_1. \end{aligned}$$

Since $\gamma_1\phi \ne 1$, the minimum of $\Delta(\gamma_2)$ is attained at $\gamma_2^* := (\gamma_1\phi+1-2\phi)\gamma_1/(\gamma_1\phi-1)^2$. Next we verify $\Delta(\gamma_2^*) \le 0$ and

$\gamma_1 \le \gamma_2^*$: We have

$$\begin{aligned}
\Delta(\gamma_2^*) \le 0 &\Leftrightarrow -\frac{(\gamma_1\phi+1-2\phi)^2\gamma_1^2}{(\gamma_1\phi-1)^2}+\gamma_1^2 \le 0\\
&\Leftrightarrow (\gamma_1\phi-1)^2 \le (\gamma_1\phi+1-2\phi)^2\\
&\Leftrightarrow (1-\phi)(\gamma_1\phi-\phi) \ge 0\\
&\Leftrightarrow (1-\phi)(\gamma_1-1) \ge 0,\\
\gamma_1 \le \gamma_2^* &\Leftrightarrow \gamma_1 \le \frac{(\gamma_1\phi+1-2\phi)\gamma_1}{(\gamma_1\phi-1)^2}\\
&\Leftrightarrow (\gamma_1\phi-1)^2 \le (\gamma_1\phi+1-2\phi)\\
&\Leftrightarrow \gamma_1^2\phi-3\gamma_1+2 \le 0,
\end{aligned}$$

and $(1-\phi)(\gamma_1-1) \ge 0$ because $\phi \in (0,1)$ and $\gamma_1 \in [1,2)$, while $\gamma_1^2\phi-3\gamma_1+2 \le 0$ holds because $\gamma_1^2\phi-3\gamma_1+2 \le (\gamma_1-1)(\gamma_1-2) \le 0$.

With the above calculations, we immediately obtain $\Delta'(\gamma_2) \le 0$ for every $\gamma_2 \in [1,\gamma_1]$. And with $\gamma_1 \ge 1$ and $\Delta(1) \ge 0$, we see ξ_1 is real-valued, $\xi_1 \ge 1$, and

$$\begin{aligned}
\gamma_1 \le \xi_1 &\Leftrightarrow \Delta(\gamma_1) \ge 0\\
&\Leftrightarrow (\gamma_1\phi-1)^2\gamma_1^2-2(\gamma_1\phi+1-2\phi)\gamma_1^2+\gamma_1^2 \ge 0\\
&\Leftrightarrow (\gamma_1\phi-1)^2-2(\gamma_1\phi+1-2\phi)+1 \ge 0\\
&\Leftrightarrow \gamma_1^2\phi-4\gamma_1+4 \ge 0\\
&\Leftrightarrow (2-\gamma_1)^2 \ge \gamma_1^2(1-\phi)\\
&\Leftrightarrow 2-\gamma_1 \ge \gamma_1\sqrt{1-\phi}\\
&\Leftrightarrow \gamma_1 \le \frac{2}{1+\sqrt{1-\phi}}.
\end{aligned}$$

We now draw the other two statements. If $\gamma_1 \in [1, \frac{2}{1+\sqrt{1-\phi}}]$, then $\gamma_1 \le \xi_1$, and therefore $\Delta(\gamma_2) \ge 0$ for every $\gamma_2 \in [1,\gamma_1]$. On the other hand, assume $\gamma_1 \in [\frac{2}{1+\sqrt{1-\phi}}, 2)$. Then $\xi_1 \le \gamma_1 \le \gamma_2^*$ and $\Delta(\gamma_1) \le 0$, so $\Delta(\gamma_2) \ge 0$ for each $\gamma_2 \in [1,\xi_1]$ and $\Delta(\gamma_2) \le 0$ for each $\gamma_2 \in [\xi_1,\gamma_1]$. □

Lemma 18. *Let $\phi \in (0,1)$ and $\gamma_1 \in [1,2)$ be fixed with $\gamma_1\phi \ne 1$. Consider the following quadratic equation in variable z and its discriminant $\Delta(\gamma_2)$:*

$$\begin{aligned}
&z^2-(2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi)z+(\gamma_1-1)(\gamma_2-1)=0,\\
&\Delta(\gamma_2) := (2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi)^2-4(\gamma_1-1)(\gamma_2-1).
\end{aligned}$$

Let $z_1(\gamma_2), z_2(\gamma_2)$ be the two roots of the above quadratic equation, and let ξ_1, ξ_2 be the two roots of $\Delta(\gamma_2)=0$ with $|\xi_1| \le |\xi_2|$. Then $1 \le \xi_1$. Moreover, with $f(\gamma_2) := \big\{|z_1(\gamma_2)|, |z_2(\gamma_2)|\big\}$, we have:

- *If $\gamma_1 \in [1, \frac{2}{1+\sqrt{1-\phi}}]$, then $f(\gamma_2)$ is non-increasing in $[1,\gamma_1]$.*
- *If $\gamma_1 \in [\frac{2}{1+\sqrt{1-\phi}}, 2)$, then $\xi_1 \le \gamma_1$, $f(\gamma_2)$ is non-increasing in $[1,\xi_1]$, and*

$$f(\gamma_2) = \sqrt{(\gamma_1-1)(\gamma_2-1)}, \quad \forall \gamma_2 \in [\xi_1,\gamma_1].$$

Proof of Lemma 18. The case with $\gamma_2 \in [\xi_1,\gamma_1]$ follows directly from Lemma 17. It remains to consider the case $\gamma_2 \in [1, \min\{\xi_1,\gamma_1\}]$, where by Lemma 17 we know $\Delta(\gamma_2) \ge 0$ and $z_1(\gamma_2), z_2(\gamma_2)$ are real-valued. Moreover, in this case we have

$$f(\gamma_2) = \frac{|2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi| + \sqrt{\Delta(\gamma_2)}}{2},$$

and therefore the derivative of $f(\gamma_2)$ is given as

$$f'(\gamma_2) = \begin{cases} \frac{-1+\gamma_1\phi}{2} + \frac{\Delta'(\gamma_2)}{4\sqrt{\Delta(\gamma_2)}} & 2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi \geq 0; \\ \frac{1-\gamma_1\phi}{2} + \frac{\Delta'(\gamma_2)}{4\sqrt{\Delta(\gamma_2)}} & 2-\gamma_1-\gamma_2+\gamma_1\gamma_2\phi < 0. \end{cases}$$

Furthermore, since $\Delta'(\gamma_2) \leq 0$ for every $\gamma_2 \in [1, \gamma_1]$ (Lemma 17), to prove $f(\gamma_2)$ is non-increasing in $\gamma_2 \in [1, \min\{\xi_1, \gamma_1\}]$ it suffices to show

$$2\sqrt{\Delta(\gamma_2)} \cdot |\gamma_1\phi - 1| \leq |\Delta'(\gamma_2)|,$$

but by Lemma 14 this indeed holds. The proof is complete. □

We are in a position to begin proving Theorem 5:

Proof of Theorem 5. We use the notations in the proof of Theorem 3, recalled here for convenience: $z_{i1}(\gamma_1, \gamma_2)$ and $z_{i2}(\gamma_1, \gamma_2)$ denote the two roots of (16) and

$$f_i(\gamma_1, \gamma_2) := \max\Big\{|z_{i1}(\gamma_1, \gamma_2)|, |z_{i2}(\gamma_1, \gamma_2)|\Big\}.$$

Define $\Delta(\gamma_1, \gamma_2)$ to be the discriminant of (16), that is

$$\Delta(\gamma_1, \gamma_2) := \Big(2-\gamma_1-\gamma_2+\gamma_1\gamma_2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)\Big)^2 - 4(\gamma_1-1)(\gamma_2-1).$$

In light of the symmetry in the spectrum of $\boldsymbol{M}(\gamma_1, \gamma_2)$ (Lemma 13), we will assume $\gamma_2 \leq \gamma_1$ without loss of generality. We proceed by considering two cases, $\gamma_1\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) \neq 1$ and $\gamma_1\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) = 1$.

Case 1: $\gamma_1\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) \neq 1$**.** For this case, we roughly follow the proof logic in *Part 2* of Theorem 3 with non-trivial modifications to handle the extra difficulty brought by optimization over S_{11}. First observe $\Delta(\gamma_1, \gamma_2)$ is quadratic in γ_2, and solving $\Delta(\gamma_1, \gamma_2) = 0$ for γ_2 gives two roots, $\xi_1(\gamma_1)$ and $\xi_2(\gamma_1)$. Without loss of generality we assume $|\xi_1| \leq |\xi_2|$. We fix $\gamma_1 \in [1, 2)$ and proceed by addressing two sub-cases, $\gamma_1 \in [1, \gamma_1^*]$ and $\gamma_1 \in [\gamma_1^*, 2)$.

Case 1.1: $\gamma_1 \in [1, \gamma_1^*]$**.** In this case, Lemma 18 suggests that $\gamma_2 \mapsto f_1(\gamma_1, \gamma_2)$ is non-increasing in $[1, \gamma_1]$. Furthermore, we have $|z_{11}(\gamma_1, \gamma_1)| \cdot |z_{12}(\gamma_1, \gamma_1)| = (\gamma_1 - 1)^2$, and therefore

$$f_1(\gamma_1, \gamma_1) \geq \gamma_1 - 1 = \gamma_2 - 1,$$

from which it follows that γ_1 is a global minimizer of

$$\min_{\gamma_2 \in [1, \gamma_1]} \Big(\max\big\{\gamma_2 - 1, f_1(\gamma_1, \gamma_2)\big\}\Big). \tag{29}$$

We next prove

$$f_i(\gamma_1, \hat{\gamma}_2) \leq \max\big\{\hat{\gamma}_2 - 1, f_1(\gamma_1, \hat{\gamma}_2)\big\}, \quad \forall i = 1, \ldots, r, \tag{30}$$

where $\hat{\gamma}_2$ is a global minimizer of (29), that is $\hat{\gamma}_2 = \gamma_1$. In other words, we will prove $f_i(\gamma_1, \gamma_1) \leq \gamma_1 - 1$ for every $i = 1, \ldots, r$. Note that $f_i(\gamma_1, \gamma_1)$ is associated with the quadratic equation

$$z^2 - \big(2 - 2\gamma_1 + \gamma_1^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\big)z + (\gamma_1 - 1)^2 = 0. \tag{31}$$

If its two roots $z_{i1}(\gamma_1, \gamma_1)$ and $z_{i2}(\gamma_1, \gamma_1)$ are complex, then $f_i(\gamma_1, \gamma_1) = \sqrt{(\gamma_1-1)^2} = \gamma_1 - 1$, so (30) holds. Then we consider the case where (31) admits real-valued roots, which means

$$\begin{aligned} \big(2 - 2\gamma_1 + \gamma_1^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\big)^2 - 4(\gamma_1-1)^2 \geq 0 &\Leftrightarrow \gamma_1^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top) \cdot \big(4 - 4\gamma_1 + \gamma_1^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\big) \geq 0 \\ &\Leftrightarrow 4 - 4\gamma_1 + \gamma_1^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top) \geq 0 \\ &\Rightarrow 2 - 2\gamma_1 + \gamma_1^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top) \geq 0. \end{aligned}$$

Here the last step follows from the fact $\gamma_1 \geq 1$. But then $2-2\gamma_1+\gamma_1^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top) \leq 2-2\gamma_1+\gamma_1^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)$, it must be the case that $f_i(\gamma_1,\gamma_1) \leq f_1(\gamma_1,\gamma_1)$, and we have proved (30). As a result, we have obtained (see, e.g., Lemma 11)

$$\gamma_1 \in \operatorname*{argmin}_{\gamma_2\in[1,\gamma_1]} \Big(\max\big\{\gamma_2-1, \max_{i=1,\ldots,r} f_i(\gamma_1,\gamma_2)\big\}\Big),$$

which, by the definition of $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$, leads to

$$\begin{aligned}
\min_{\substack{\gamma_1\in[1,\gamma_1^*]\\ \gamma_2\in[1,\gamma_1]}} \rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big) &= \min_{\gamma_1\in[1,\gamma_1^*]} \max\Big\{\gamma_1-1, \min_{\gamma_2\in[1,\gamma_1]} \Big(\max\big\{\gamma_2-1, \max_{i=1,\ldots,r} f_i(\gamma_1,\gamma_2)\big\}\Big)\Big\} \\
&= \min_{\gamma_1\in[1,\gamma_1^*]} \max\Big\{\gamma_1-1, \max_{i=1,\ldots,r} f_i(\gamma_1,\gamma_1)\Big\} \\
&= \min_{\gamma_1\in[1,\gamma_1^*]} \rho\big(\boldsymbol{M}(\gamma_1,\gamma_1)\big).
\end{aligned}$$

We can now finish the proof for *Case 1.1* by invoking Proposition 2.

Case 1.2: $\gamma_1 \in [\gamma_1^*, 2)$**.** Similarly to *Case 1.1*, from Lemma 18, we obtain $1 \leq \xi_1(\gamma_1) \leq \gamma_1$ and $\gamma_2 \mapsto f_1(\gamma_1,\gamma_2)$ is non-increasing in $[1,\xi_1(\gamma_1)]$. In particular, we have $\Delta(\gamma_1,\xi_1(\gamma_1))=0$, which means $z_{11}\big(\gamma_1,\xi_1(\gamma_1)\big) = z_{12}\big(\gamma_1,\xi_1(\gamma_1)\big)$ and therefore

$$f_1\big(\gamma_1,\xi_1(\gamma_1)\big) = \sqrt{(\gamma_1-1)\big(\xi_1(\gamma_1)-1\big)} \geq \xi_1(\gamma_1)-1.$$

Lemma 18 implies $\Delta(\gamma_1,\gamma_2)\leq 0$ for every $\gamma_2 \in [\xi_1(\gamma_1),\gamma_1]$, so $f(\gamma_1,\gamma_2) = \sqrt{(\gamma_1-1)(\gamma_2-1)}$. We can now conclude $\xi_1(\gamma_1)$ is a global minimizer of (29), and moreover, we have

$$\begin{aligned}
\min_{\substack{\gamma_1\in[\gamma_1^*,2)\\ \gamma_2\in[1,\gamma_1]}} \max\Big\{\gamma_1-1,\gamma_2-1,f_1(\gamma_1,\gamma_2)\Big\} &= \min_{\gamma_1\in[\gamma_1^*,2)} \max\Big\{\gamma_1-1, \sqrt{(\gamma_1-1)\big(\xi_1(\gamma_1)-1\big)}\Big\} \\
&= \gamma_1^*-1,
\end{aligned}$$

where the last equality follows from the fact $\xi(\gamma_1^*)\leq\gamma_1^*$, proved in Lemma 18. We have thus obtained a lower bound of $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2\big)$, that is

$$\begin{aligned}
\rho\big(\boldsymbol{M}(\gamma_1^*,\gamma_1^*)\big) &= \gamma_1^*-1 \\
&= \min_{\gamma_1\in[\gamma_1^*,2),\gamma_2\in[1,\gamma_1]} \max\Big\{\gamma_1-1,\gamma_2-1,f_1(\gamma_1,\gamma_2)\Big\} \\
&\leq \min_{\gamma_1\in[\gamma_1^*,2),\gamma_2\in[1,\gamma_1]} \max\Big\{\gamma_1-1,\gamma_2-1,\max_{i=1,\ldots,r} f_i(\gamma_1,\gamma_2)\Big\} \\
&= \min_{\gamma_1\in[\gamma_1^*,2),\gamma_2\in[1,\gamma_1]} \rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big).
\end{aligned}$$

We need to prove the above inequality is actually an equality. To do so, we can show $f_i(\gamma_1^*,\gamma_1^*)\leq\gamma_1^*-1$ for every $i=1,\ldots,r$. Note that $f_i\big(\gamma_1^*,\gamma_1^*\big)$ is associated with the equation

$$z^2-\big(2-2\gamma_1^*+(\gamma_1^*)^2\cdot\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\big)z+(\gamma_1^*-1)^2=0,$$

but the definition of γ_1^* implies $4-4\gamma_1^*+(\gamma_1^*)^2\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)=0$ and therefore

$$4-4\gamma_1^*+(\gamma_1^*)^2\lambda_i(\boldsymbol{C}\boldsymbol{C}^\top)\leq 0, \quad \forall i=1,\ldots,r.$$

This means $f_i(\gamma_1^*,\gamma_1^*)=\gamma_1^*-1$ for every $i=1,\ldots,r$. We finished *Case 1*.

Case 2: $\gamma_1\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)=1$**.** We will show that the minimum of $\rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big)$ in this case is larger than or equal to its minimum γ_1^*-1 in *Case 1*, that is,

$$\min_{\gamma_1\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)=1,\gamma_2\in[1,\gamma_1]} \rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big) \geq \gamma_1^*-1. \tag{32}$$

We consider two subcases, $\gamma_1 \geq 4/3$ and $\gamma_1 < 4/3$.

Case 2.1: $\gamma_1 \geq 4/3$**.** First, recall $\gamma_1^* = \frac{2}{1+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}$ and note that

$$\begin{aligned}\gamma_1 \geq \frac{4}{3} &\Leftrightarrow \lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) \leq \frac{3}{4}\\ &\Leftrightarrow 2\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)} \geq 1\\ &\Leftrightarrow 1 \geq 2\left(1-\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}\right)\\ &\Leftrightarrow \frac{1}{\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)} \geq \frac{2}{1+\sqrt{1-\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top)}}\\ &\Leftrightarrow \gamma_1 - 1 \geq \gamma_1^* - 1.\end{aligned}$$

But we know from Lemma 13 that $\gamma_1 - 1$ is an eigenvalue of $\boldsymbol{M}(\gamma_1,\gamma_2)$, therefore (32) holds.

Case 2.2: $\gamma_1 < 4/3$**.** Note that $1 \leq \gamma_2 \leq \gamma_1 < 2$, so we can write

$$\begin{aligned}\gamma_1 < \frac{4}{3} &\Rightarrow 2-\gamma_1 > 2(\gamma_1-1)\\ &\Rightarrow (2-\gamma_1)^2 > 4(\gamma_1-1)^2\\ &\Rightarrow (2-\gamma_1)^2 > 4(\gamma_1-1)(\gamma_2-1).\end{aligned}$$

This means the quadratic equation

$$z^2 - (2-\gamma_1)z + (\gamma_1-1)(\gamma_2-1) = 0$$

always has two different roots. Moreover, since $\gamma_1\lambda_1(\boldsymbol{C}\boldsymbol{C}^\top) = 1$, Lemma 13 implies these two roots are eigenvalues of $\boldsymbol{M}(\gamma_1,\gamma_2)$, and this quadratic equation coincides with (16), whose roots were denoted previously by $z_{11}(\gamma_1,\gamma_2)$ and $z_{12}(\gamma_1,\gamma_2)$. We can then write

$$\begin{aligned}f_1(\gamma_1,\gamma_2) &= \max\{z_{11}(\gamma_1,\gamma_2), z_{12}(\gamma_1,\gamma_2)\}\\ &= \frac{2-\gamma_1}{2} + \frac{\sqrt{(2-\gamma_1)^2 - 4(\gamma_1-1)(\gamma_2-1)}}{2}.\end{aligned}$$

Therefore $\gamma_2 \mapsto f_1(\gamma_1,\gamma_2)$ is decreasing in $[1,\gamma_1]$, and we have thus obtained

$$f(\gamma_1,\gamma_1) = \min_{\gamma_2\in[1,\gamma_1]} f_1(\gamma_1,\gamma_2).$$

We finish the proof by observing

$$\min_{\gamma_2\in[1,\gamma_1]} \rho\big(\boldsymbol{M}(\gamma_1,\gamma_2)\big) \geq \max\left\{\gamma_1 - 1, f_1(\gamma_1,\gamma_1)\right\} \geq \gamma_1^* - 1,$$

where the last ineequality follows from Proposition 2 with the special case $r = 1$. □

D. Extra Discussions on Gauss–Seidel

The classic Gauss-Seidel method applied to normal equations $\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^\top\boldsymbol{y}$ is precisely a BGD method applied to (1) with n blocks, each of size 1, and block i is with stepsize $1/a_{ii}$, where a_{ii} is the i-th diagonal entry of $\boldsymbol{A}^\top\boldsymbol{A}$. This stepsize rule also corresponds to block exact minimization. Again, this choice of stepsizes is sub-optimal. To see this, consider a simple case where $\boldsymbol{A}$ has only two columns $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ ($n = 2$). By Lemma 1, the iterates of BGD satisfy $\boldsymbol{x}^+ - \boldsymbol{x}^* = \boldsymbol{M}(\gamma_1,\gamma_2)\cdot(\boldsymbol{x}-\boldsymbol{x}^*)$ with $\boldsymbol{M}(\gamma_1,\gamma_2)$ defined as

$$\boldsymbol{M}(\gamma_1,\gamma_2) := \begin{bmatrix} 1-\gamma_1\boldsymbol{a}_1^\top\boldsymbol{a}_1 & -\gamma_1\boldsymbol{a}_1^\top\boldsymbol{a}_2 \\ -\gamma_2(1-\gamma_1\boldsymbol{a}_1^\top\boldsymbol{a}_1)\boldsymbol{a}_2^\top\boldsymbol{a}_1 & \gamma_1\gamma_2(\boldsymbol{a}_1^\top\boldsymbol{a}_2)^2 + 1 - \gamma_2\boldsymbol{a}_2^\top\boldsymbol{a}_2 \end{bmatrix}.$$

The Gauss-Seidel method takes stepsizes $\gamma_1 = 1/(\boldsymbol{a}_1^\top \boldsymbol{a}_1), \gamma_2 = 1/(\boldsymbol{a}_2^\top \boldsymbol{a}_2)$, and one verifies that

$$\rho\bigg(\boldsymbol{M}\Big(\frac{1}{\boldsymbol{a}_1^\top \boldsymbol{a}_1}, \frac{1}{\boldsymbol{a}_2^\top \boldsymbol{a}_2}\Big)\bigg) = (\boldsymbol{a}_1^\top \boldsymbol{a}_2)^2/\big((\boldsymbol{a}_1^\top \boldsymbol{a}_1)\cdot(\boldsymbol{a}_2^\top \boldsymbol{a}_2)\big).$$

On the other hand, it follows from Remark 7 that the minimum spectral radius in this case is zero; indeed, with stepsizes $\gamma_1 = \frac{1}{\boldsymbol{a}_1^\top \boldsymbol{a}_1}, \gamma_2 = \frac{\boldsymbol{a}_1^\top \boldsymbol{a}_1}{(\boldsymbol{a}_1^\top \boldsymbol{a}_1)\cdot(\boldsymbol{a}_2^\top \boldsymbol{a}_2)-(\boldsymbol{a}_1^\top \boldsymbol{a}_2)^2}$ we have

$$\boldsymbol{M}(\gamma_1, \gamma_2) = \begin{bmatrix} 0 & -\boldsymbol{a}_1^\top \boldsymbol{a}_2/(\boldsymbol{a}_1^\top \boldsymbol{a}_1) \\ 0 & 0 \end{bmatrix},$$

whose eigenvalues are zero. Note that this reasoning only provides a counter-example showing the stepsize sub-optimality of the Gauss-Seidel method, and we are none the wiser: Finding closed-form optimal stepsizes of BGD in its full generality has remained non-trivial for $n > 2$.

E. Proof for Theorems Related to Generalized Alternating Projection

Proof of Lemma 5. It follows directly from Exercise 2.8 of Vidal et al. (2016), and either Property 2.1 of Zhu & Knyazev (2013) or Theorem 2.7 of Knyazev & Argentati (2007). □

Proof of Lemma 6. With $C := A_2^\top A_1$, we have

$$\begin{aligned} A^\top A &= \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix}, \\ A_2^\top A &= \begin{bmatrix} C & \boldsymbol{I}_2 \end{bmatrix}, \\ A_1^\top A &= \begin{bmatrix} \boldsymbol{I}_1 & C^\top \end{bmatrix}. \end{aligned}$$

we need to prove

$$A^\top(\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)A = \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix}\left(\boldsymbol{I} - \begin{bmatrix} 0 & 0 \\ \gamma_2 C & \gamma_2 \boldsymbol{I}_2 \end{bmatrix}\right)\left(\boldsymbol{I} - \begin{bmatrix} \gamma_1 \boldsymbol{I}_1 & \gamma_1 C^\top \\ 0 & 0 \end{bmatrix}\right).$$

We do so by first simplifying the two terms separately. The left-hand side can be written as

$$\begin{aligned} A^\top(\boldsymbol{I} - \gamma_2 A_2 A_2^\top)(\boldsymbol{I} - \gamma_1 A_1 A_1^\top)A &= A^\top A - \gamma_2 A^\top A_2 A_2^\top A - \gamma_1 A^\top A_1 A_1^\top A + \gamma_1\gamma_2 A^\top A_2 C A_1^\top A \\ &= \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix} - \gamma_2 \begin{bmatrix} C^\top C & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix} - \gamma_1 \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & CC^\top \end{bmatrix} + \gamma_1\gamma_2 \begin{bmatrix} C^\top \\ \boldsymbol{I}_2 \end{bmatrix} C \begin{bmatrix} \boldsymbol{I}_1 & C^\top \end{bmatrix}, \end{aligned}$$

and the right-hand side can be written as

$$\begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix} - \gamma_2 \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ C & \boldsymbol{I}_2 \end{bmatrix} - \gamma_1 \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix}\begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ 0 & 0 \end{bmatrix} + \gamma_1\gamma_2 \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ C & CC^\top \end{bmatrix}.$$

Then, by inspection, it suffices to prove the following three matrix equations:

$$\begin{aligned} \begin{bmatrix} C^\top C & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix} &= \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ C & \boldsymbol{I}_2 \end{bmatrix}, \\ \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & CC^\top \end{bmatrix} &= \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix}\begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ 0 & 0 \end{bmatrix}, \\ \begin{bmatrix} C^\top \\ \boldsymbol{I}_2 \end{bmatrix} C \begin{bmatrix} \boldsymbol{I}_1 & C^\top \end{bmatrix} &= \begin{bmatrix} \boldsymbol{I}_1 & C^\top \\ C & \boldsymbol{I}_2 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ C & CC^\top \end{bmatrix}. \end{aligned}$$

These can be easily verified, therefore the proof is complete. □

Proof of Corollary 1. By Theorem 2 we have

$$\min_{\gamma_1>0,\gamma_2>0}\rho\big(M(\gamma_1,\gamma_2)\big)=\begin{cases}\frac{\sqrt{1-\lambda_r(CC^\top)}-\sqrt{1-\lambda_1(CC^\top)}}{\sqrt{1-\lambda_r(CC^\top)}+\sqrt{1-\lambda_1(CC^\top)}} & r=\min\{n_1,n_2\};\\ \frac{1-\sqrt{1-\lambda_1(CC^\top)}}{1+\sqrt{1-\lambda_1(CC^\top)}}. & r<\min\{n_1,n_2\}.\end{cases}$$

By Lemma 5 we have $\sqrt{1-\lambda_r(CC^\top)}=\sin(\theta_r)$ and $\sqrt{1-\lambda_1(CC^\top)}=\sin(\theta_1)$. The proof is then complete. □

Remark 12. Here we describe the corresponding stepsizes γ_1^*,γ_2^* that attain the above minimum. This is derived from Theorem 5 and the proof of Theorem 4.

- If $r=\min\{n_1,n_2\}$, then we have

$$\gamma_1^*=\left(\frac{\sqrt{(1+\sin(\theta_1))(1+\sin(\theta_r))}+\sqrt{(1-\sin(\theta_1))(1-\sin(\theta_r))}}{\sin(\theta_r)+\sin(\theta_1)}\right)^2,$$
$$\gamma_2^*=\left(\frac{\sqrt{(1+\sin(\theta_1))(1+\sin(\theta_r))}-\sqrt{(1-\sin(\theta_1))(1-\sin(\theta_r))}}{\sin(\theta_r)+\sin(\theta_1)}\right)^2,$$

  or

$$\gamma_1^*=\left(\frac{\sqrt{(1+\sin(\theta_1))(1+\sin(\theta_r))}-\sqrt{(1-\sin(\theta_1))(1-\sin(\theta_r))}}{\sin(\theta_r)+\sin(\theta_1)}\right)^2,$$
$$\gamma_2^*=\left(\frac{\sqrt{(1+\sin(\theta_1))(1+\sin(\theta_r))}+\sqrt{(1-\sin(\theta_1))(1-\sin(\theta_r))}}{\sin(\theta_r)+\sin(\theta_1)}\right)^2.$$

  In other words, in this case there are two sets of optimal parameters and they are symmetric. These stepsizes would coincide with the stepsize rule of Fält & Giselsson (2017) if θ_r were equal to $\pi/2$. As justified, for generic subspaces we have $\theta_r=\pi/2$ with probability 0.

- if $r<\min\{n_1,n_2\}$, then

$$\gamma_1^*=\gamma_2^*=\frac{2}{1+\sin(\theta_1)},$$

  which coincides with the stepsize rule of Fält & Giselsson (2017); see Table 1.

F. Code

F.1. Experiments on Least-Squares With Assumption 1 (Figure 1)

```
clc; clear all; format longG

cond_num = 1e5;
n_iter = 5000;

num_trials = 1;
m = 1000;
n1 = 300;
n2 = 500;
n = n1 + n2;
noise_level = 0.01;

errorsBGD = zeros(num_trials, n_iter+1);
errorsGD = zeros(num_trials, n_iter+1);
errorsHB = zeros(num_trials, n_iter+1);

times_BGD = zeros(num_trials, 1);
```

```
% stepsizes of methods
global gammaGD alphaHB betaHB gamma1 gamma2;

for t=1:num_trials
    %% generate data
    [A, y, x, lambdamaxCC, lambdaminCC] = gen_data(m, n1, n2, noise_level, cond_num);

    %% calculae stepsizes
    minC = sqrt(1-lambdamaxCC);
    maxC = sqrt(1-lambdaminCC);

    lambdamaxAA = 1 + sqrt(lambdamaxCC);
    lambdaminAA = 1 - sqrt(lambdamaxCC);

    lambdamaxAA/lambdaminAA

    gammaGD = 2 / (lambdamaxAA + lambdaminAA);

    alphaHB = 4 / ( sqrt(lambdamaxAA) + sqrt(lambdaminAA) )^2;
    betaHB = ( sqrt(lambdamaxAA) - sqrt(lambdaminAA) )^2;
    betaHB = betaHB / ( sqrt(lambdamaxAA) + sqrt(lambdaminAA) )^2;


    gamma1 = (1+maxC)*(1+minC) / (maxC + minC)^2 ;
    gamma2 = gamma1;

    %% run algorithm
    [BGD_x, BGD_iters, BGD_time] = BGD(A, y, n1, n2, n_iter);
    errorsBGD(t,:) = vecnorm(BGD_iters - x, 2, 1);

    [HB_x, HB_iters, HB_time] = HB(A, y, n_iter);
    errorsHB(t,:) = vecnorm(HB_iters - x, 2, 1);

    [GD_x, GD_iterates, GD_time] = GD(A, y, n_iter);
    errorsGD(t,:) = vecnorm(GD_iterates - x, 2, 1);
end

mean_errorsBGD = mean(errorsBGD,1);
mean_errorsGD = mean(errorsGD,1);
mean_errorsHB = mean(errorsHB,1);


[mean_errorsBGD; mean_errorsHB; mean_errorsGD]

function [BGD_x, BGD_iters, BGD_time] = BGD(A, y, n1, n2, n_iter)
    tic;

    global gamma1 gamma2;

    c1 = 1-gamma1; c2 = 1-gamma2;

    A1 = A(:, 1:n1);
    A2 = A(:, n1+1:end);

    C = A2'*A1;
    Ay = A'*y;


    BGD_x = zeros(n1+n2,1);
    BGD_iters = BGD_x;
    for i=1:n_iter
        BGD_x(1:n1) = c1*BGD_x(1:n1) - gamma1*(C'*BGD_x(n1+1:end) - Ay(1:n1));
        BGD_x(n1+1:end) = c2*BGD_x(n1+1:end) - gamma2*(C*BGD_x(1:n1) - Ay(n1+1:end));
```

```
        BGD_iters = [BGD_iters BGD_x];
    end
    BGD_time = toc;
end

function [GD_x, GD_iterates, GD_time] = GD(A, y, n_iter)
    tic;

    global gammaGD;

    [m,n] = size(A);

    AA = A'*A;
    Ay = A'*y;

    GD_x = zeros(n,1);
    GD_iterates = GD_x;
    for i=1:n_iter
        GD_x = GD_x - gammaGD * (AA*GD_x - Ay);

        GD_iterates = [GD_iterates GD_x];
    end

    GD_time = toc;
end

function [HB_x, HB_iters, HB_times] = HB(A, y, n_iter)
    tic;

    global alphaHB betaHB;

    [m,n] = size(A);

    AA = A'*A;
    Ay = A'*y;

    HB_x = zeros(n,1); HB_x_old = HB_x;
    HB_iters = HB_x;
    HB_times = [];
    for i=1:n_iter
        HB_x_new = HB_x - alphaHB * (AA*HB_x - Ay) + betaHB*(HB_x - HB_x_old);

        HB_x_old = HB_x; HB_x = HB_x_new;

        HB_iters = [HB_iters HB_x];
        HB_times = [HB_times toc];
    end

end


function [A, y, x, lambdamaxCC, lambdaminCC]= gen_data(m, n1, n2, noise_level, cond_num)
    % cond_num = ( 1+sqrt(lambdamaxC) ) / ( 1-sqrt(lambdamaxC) )

    lambdamaxC = (cond_num - 1) / (cond_num + 1);

    %% generate

    [C, lambdamaxCC, lambdaminCC] = gen_mat_random_bounded_svs(n2, n1, lambdamaxC);

    [U, ~, ~] = svd(randn(m-n2, n1), 'econ');

    scale = vecnorm(C, 2, 1);
    U = U .* sqrt(1- scale.^2);
```

```
    A1 = [C; U];              % by construction, A1 is orthogonal (A1'*A1 = identity)
    A2 = [eye(n2,n2); zeros(m-n2, n2)];

    [U1, ~, ~] = svd(randn(n1,n1)); A1 = A1 * U1;
    [U2, ~, ~] = svd(randn(n2,n2)); A2 = A2 * U2;

    A = [A1 A2];

    x = randn(n1+n2, 1);

    noise = randn(m,1); noise = noise / norm(noise);

    y = A*x + noise_level* noise;

    x = A \ y;
end

function [C, lambdamaxCC, lambdaminCC] = gen_mat_random_bounded_svs(n2, n1, L)
    % generate a n2xn1 matrix C such that
    % the maximum and minimum singlar value of C are U and L respectively
    %    the rest singular values are randomly chosen in [0,L2]

    num = min(n2,n1) - 1;

    svs = L*rand(num,1);


    svs = sort([svs; L], 'descend');

    lambdamaxCC = max(svs)*max(svs);
    lambdaminCC = min(svs)*min(svs);


    C = diag(svs);

    if n1 > n2
        C = [C zeros(n2, n1-n2)];
    elseif n1 < n2
        C = [C; zeros(n2-n1, n1)];
    end
end
```

F.2. Experiments on Least-Squares (Without Assumption 1)

Below is the code implementing the heavy ball method in comparison to our proposal, block-wise QR orthogonalization followed by BGD with the optimal stepsizes. As analyzed in the main paper, the code makes the point that the proposed method would converge faster than the heavy ball method under the current problem configuration. That being said, it should also be noted that this is still slower than practical least-squares solvers, e.g., QR orthogonalization followed by back substitution to solve an upper triangular system, randomized methods, (preconditioned) conjugate gradient methods, or the MATLAB backslash solver.

```
clc; clear all; format longG

%% setup
cond_num = 1e5;
n_iter = 5000;

num_trials = 1;
m = 1000;
n1 = 300;
n2 = 500;
n = n1 + n2;
```

```
noise_level = 0.01;


%% run experiments
errorsBGD = zeros(num_trials, n_iter+1);
errorsHB = zeros(num_trials, n_iter+1);

for t=1:num_trials
    %% generate data
    [A, y, x] = gen_data(m, n, cond_num, noise_level);
    %% calculate stepsizes

    %% run algorithm
    [BGD_x, Q1, Q2, R1, R2, BGD_iters, qrtime, BGD_times] = BGD(A, y, n1, n2, n_iter);
    % the LS solution [x1, x2] can be found by backward substitution using R1 and R2.

    errorsBGD(t,:) = Q1 * BGD_iters(1:n1, :) + Q2 * BGD_iters(n1+1:end, :) - y
    errorsBGD(t,:) = vecnorm(errorsBGD(t,:), 2, 1);

    [HB_x, HB_iters, HB_times] = HB(A, y, n_iter);
    % errorsHB(t,:) = vecnorm(HB_iters - x, 2, 1);
    errorsHB(t,:) = vecnorm(A*HB_iters - y, 2, 1);
end

mean_errorsBGD = mean(errorsBGD,1);
mean_errorsHB = mean(errorsHB,1);

BGD_qrtimes = qrtime + BGD_times;

figure(1);
plot(BGD_times, 'LineWidth', 2);
hold on;
plot(BGD_times + qrtime, 'LineWidth', 2);

plot(HB_times, 'LineWidth', 2);
legend("BGD", "BGD+Ortho", "HB")

set(gca, 'FontSize', 18)

figure(2);
plot(mean_errorsBGD, 'LineWidth', 2);
hold on;

plot(mean_errorsHB, 'LineWidth', 2);
legend("BGD+Ortho", "HB")

set(gca, 'YScale', 'log')
set(gca, 'FontSize', 18)



function [BGD_x, Q1, Q2, R1, R2, BGD_iters, qrtime, BGD_times] = BGD(A, y, n1, n2, n_iter)

    tic; % orthogonalize
    [Q1, R1] = qr(A(:, 1:n1), 0);
    [Q2, R2] = qr(A(:, n1+1:end), 0);
    qrtime = toc;

    % calculate stepsizes. We don't count the running times here.
    C = Q2'*Q1;

    s = svd(C, 0);

    gamma1 = 2/(1 + sqrt(1-max(s)^2)); gamma2 = gamma1;
```

```
    tic;  % run the algorithm
    c1 = 1-gamma1; c2 = 1-gamma2;

    Qy = [Q1 Q2]'*y;

    BGD_x = zeros(n1+n2,1);
    BGD_iters = BGD_x;
    BGD_times = [];
    for i=1:n_iter

        BGD_x(1:n1) = c1*BGD_x(1:n1) - gamma1*(C'*BGD_x(n1+1:end) - Qy(1:n1));
        BGD_x(n1+1:end) = c2*BGD_x(n1+1:end) - gamma2*(C*BGD_x(1:n1) - Qy(n1+1:end));

        BGD_iters = [BGD_iters BGD_x];
        BGD_times = [BGD_times toc];
    end
end

function [HB_x, HB_iters, HB_times] = HB(A, y, n_iter)
    % calculate stepsizes for HB. We don't count the running times here.
    e = eig(A'*A);
    lambdamaxAA = max(e);
    lambdaminAA = min(e);

    alphaHB = 4 / ( sqrt(lambdamaxAA) + sqrt(lambdaminAA) )^2;
    betaHB = ( sqrt(lambdamaxAA) - sqrt(lambdaminAA) )^2;
    betaHB = betaHB / ( sqrt(lambdamaxAA) + sqrt(lambdaminAA) )^2;

    tic; % run the algorithm
    [m,n] = size(A);

    AA = A'*A;
    Ay = A'*y;

    HB_x = zeros(n,1); HB_x_old = HB_x;
    HB_iters = HB_x;
    HB_times = [];
    for i=1:n_iter

        HB_x_new = HB_x - alphaHB * (AA*HB_x - Ay) + betaHB*(HB_x - HB_x_old);


        HB_x_old = HB_x; HB_x = HB_x_new;


        HB_iters = [HB_iters HB_x];
        HB_times = [HB_times toc];
    end

end

function [A, y, x] = gen_data(m, n, cond_num, noise_level)

    A = randn(m,n);

    cond_numA = sqrt(cond_num);

    [U, ~, V] = svd(A, 'econ');
    s = 1 + (cond_numA - 1).*rand(n-2,1);


    s = sort([s; 1; cond_numA], 'descend');

    A = U * diag(s) * V';
```

```
    x = randn(n,1);

    noise = randn(m,1); noise = noise / norm(noise);

    y = A*x + noise_level* noise; % just to make sure that the minimum loss is not zero

    x = A \ y;
end
```