THE APPEALS OF QUADRATIC MAJORIZATION-MINIMIZATION

Marc Robini, Lihui Wang, Yue-Min Zhu

Journal of Global Optimization

The Appeals of Quadratic Majorization-Minimization

--Manuscript Draft--

Manuscript Number:	JOGO-D-22-00253R1
Full Title:	The Appeals of Quadratic Majorization-Minimization
Article Type:	Manuscript
Keywords:	Differentiable optimization; majorization-minimization; tame optimization; multidimensional scaling; inverse problems
Corresponding Author:	Marc Robini, Ph.D. CREATIS, INSA-Lyon Villeurbanne, FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	CREATIS, INSA-Lyon
Corresponding Author's Secondary Institution:	
First Author:	Marc Robini, Ph.D.
First Author Secondary Information:	
Order of Authors:	Marc Robini, Ph.D. Lihui Wang Yuemin Zhu
Order of Authors Secondary Information:	
Funding Information:	National Natural Science Foundation of China (62161004) — Prof. Lihui Wang Science and Technology Program of Guizhou Province (Qiankehe ZK[2021]002) — Prof. Lihui Wang Région Auvergne-Rhône-Alpes (PAI 2000688501-40890) — Prof. Yuemin Zhu
Abstract:	Majorization-minimization (MM) is a versatile optimization technique that operates on surrogate functions satisfying tangency and domination conditions. Our focus is on differentiable optimization using inexact MM with quadratic surrogates, which amounts to approximately solving a sequence of symmetric positive definite systems. We begin by investigating the convergence properties of this process, from subconvergence to R-linear convergence, with emphasis on tame objectives. Then we provide a numerically stable implementation based on truncated conjugate gradient. Applications to multidimensional scaling and regularized inversion are discussed and illustrated through numerical experiments on graph layout and X-ray tomography. In the end, quadratic MM not only offers solid guarantees of convergence and stability, but is robust to the choice of its control parameters.
Response to Reviewers:	We gratefully thank the editors and the referee for examining our work. Our answers to the referee are given in the file "revision_report.pdf". Changes and additions to the manuscript are displayed in red or highlighted in yellow in the file "revised_manuscript.pdf". Best regards, Marc Robini

THE APPEALS OF QUADRATIC MAJORIZATION-MINIMIZATION

MARC C. ROBINI, LIHUI WANG, AND YUEMIN ZHU

Abstract. Majorization-minimization (MM) is a versatile optimization technique that operates on surrogate functions satisfying tangency and domination conditions. Our focus is on differentiable optimization using inexact MM with quadratic surrogates, which amounts to approximately solving a sequence of symmetric positive definite systems. We begin by investigating the convergence properties of this process, from subconvergence to R-linear convergence, with emphasis on tame objectives. Then we provide a numerically stable implementation based on truncated conjugate gradient. Applications to multidimensional scaling and regularized inversion are discussed and illustrated through numerical experiments on graph layout and X-ray tomography. In the end, quadratic MM not only offers solid guarantees of convergence and stability, but is robust to the choice of its control parameters.

Contents

Date: August 11, 2023.
2010 *Mathematics Subject Classification.* 65K05, 68W40, 90C26.
Key words and phrases. Differentiable optimization, majorization-minimization, tame optimization, multidimensional scaling, inverse problems.
This work was supported by the National Natural Science Foundation of China under grant 62161004, the Guizhou Science and Technology plan project under grant Qiankehe ZK[2021]002, the 2018 PHC-Cai Yuanpei program under grant 41400TC, the Auvergne-Rhône-Alpes region (PAI project no. 2000688501-40890), and the CNRS international research project METISLAB..
The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

1. Introduction

Majorization-minimization (MM) is a general principle for designing descent algorithms [1–3]. An MM iteration for minimizing an objective $f : \mathbb{R}^n \to \mathbb{R}$ consists of two steps. First, majorize f by a simpler surrogate function tangent to f at the current iterate. Second, minimize the surrogate function to obtain the next iterate. More precisely, a surrogate for f is a family of functions $f(\cdot\,|\,\boldsymbol{x})$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$f(\boldsymbol{x}\,|\,\boldsymbol{x}) = f(\boldsymbol{x}) \quad \text{and} \quad f(\boldsymbol{y}\,|\,\boldsymbol{x}) \geqslant f(\boldsymbol{y}). \tag{1.1}$$

Starting with an initial guess $\boldsymbol{x}_0$, an (exact) MM algorithm generates a sequence of iterates $\boldsymbol{x}_p$ by taking

$$\boldsymbol{x}_{p+1} \in \operatorname*{arg\,min}_{\boldsymbol{y} \in \mathbb{R}^n} f(\boldsymbol{y}\,|\,\boldsymbol{x}_p) =: \boldsymbol{\Phi}_0(\boldsymbol{x}). \tag{1.2}$$

This process goes downhill, since

$$f(\boldsymbol{x}_{p+1}) \leqslant f(\boldsymbol{x}_{p+1}\,|\,\boldsymbol{x}_p) \leqslant f(\boldsymbol{x}_p\,|\,\boldsymbol{x}_p) = f(\boldsymbol{x}_p). \tag{1.3}$$

It is usually assumed that f is continuous and coercive and that the map $\boldsymbol{\Phi}_0$ is single-valued and continuous, which ensures that the sequence $\{\boldsymbol{x}_p\}$ converges to the set of fixed points of $\boldsymbol{\Phi}_0$ (see, e.g., [3, Proposition 7.3.2]). If, in addition, f and all the surrogate functions $f(\cdot\,|\,\boldsymbol{x})$ are continuously differentiable, then $\{\boldsymbol{x}_p\}$ converges to the set $\mathcal{S}_f$ of stationary points of f (see [4] and the references therein), and hence converges in norm to a stationary point when $\mathcal{S}_f$ is discrete.

We focus on minimizing differentiable objectives using MM with quadratic surrogate functions. This framework, which we call *quadratic majorization-minimization* (QMM), has two advantages. First, it allows to refine the convergence results for MM and to derive weak conditions for finite-length convergence (that is, $\sum \|\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\| < \infty$). Second, QMM is well suited to large-scale problems in that it can be efficiently implemented using conjugate gradient (CG) to minimize the surrogate functions. Applications of QMM include logistic regression [5], compressed sensing [6], phase retrieval [7], target tracking in sensor networks [8], and the general problems of multidimensional scaling [9] and regularized inversion [10].

It is important to realize that there is no need to minimize the surrogate function for the descent property (1.3) to hold: it suffices that $f(\cdot\,|\,\boldsymbol{x}_p)$ be decreased at each iteration. This suggests that MM is stable even when $\boldsymbol{x}_{p+1}$ approximates the minimizer of $f(\cdot\,|\,\boldsymbol{x}_p)$, making it of great practical interest. To account for inexactness in the minimization step, we consider the iteration

$$\boldsymbol{x}_{p+1} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x}_p), \quad \gamma \in (0, 1), \tag{1.4}$$

where $\boldsymbol{\Phi}_\gamma(\boldsymbol{x})$ is the sublevel set of $f(\cdot\,|\,\boldsymbol{x})$ at height $(1-\gamma)\min f(\cdot\,|\,\boldsymbol{x}) + \gamma f(\boldsymbol{x})$, as illustrated in Figure 1. We call the constant γ the *contraction number*.

Our contribution is threefold. First, we investigate the different types of convergence of inexact QMM, from weak to strong: subconvergence, convergence in norm, finite-length convergence, R-sublinear convergence (meaning that there are a point $\boldsymbol{x}$ and an integer $s \geqslant 1$ such that $\|\boldsymbol{x}_p - \boldsymbol{x}\| = O(p^{-1/s})$), and R-linear convergence (that is, $\|\boldsymbol{x}_p - \boldsymbol{x}\| = O(\eta^p)$ for some $\eta \in (0, 1)$). Second, we provide a large-scale implementation using truncated CG, with emphasis on numerical stability. Third, we describe applications to multidimensional scaling and regularized inversion, and we present numerical experiments illustrating the convergence results.

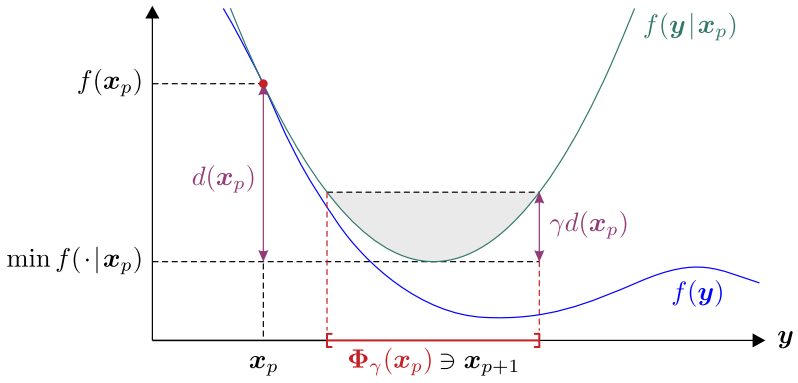

FIGURE 1. Inexact QMM iteration: $\boldsymbol{x}_{p+1}$ belongs to the sub-level set of the quadratic surrogate function $f(\cdot|\boldsymbol{x}_p)$ at height $\min f(\cdot|\boldsymbol{x}_p) + \gamma d(\boldsymbol{x}_p)$.

Note that MM procedures have also been studied in the general context of nonsmooth optimization [11–16], with the following limitations (which will be described in more detail later): on the theoretical side, the convergence guarantees are weak and/or do not cover all differentiable objectives, not even tame ones; on the implementation side, the question of the feasibility of the inner optimization problems is not addressed, and nor are the computational issues in minimizing the surrogate functions (let alone numerical stability). Our results complement those for nonsmooth MM in all these respects.

The outline of this paper is as follows. After introducing some notation in Section 2, inexact QMM algorithms are described in Section 3, followed by elementary properties in Section 4. The exposition is then structured in two parts.

The first part (Sections 5–11) is about theoretical convergence. We begin with subconvergence in Section 5, which is the basis for the global and local convergence results established in Sections 6 and 7. In Section 8 we exploit the ubiquitous Kurdyka-Łojasiewicz (KL) inequality to show that QMM sequences form trajectories of finite length. Sections 9 and 10 are devoted to two special classes of objectives that together cover most applications of differentiable optimization: tame subanalytic $\mathscr{C}^1$ objectives, for which the rate of convergence is R-linear or -sublinear depending on the flatness around the limit of the sequence; and $\mathscr{C}^2$ objectives, for which convergence is R-linear when the limit $\boldsymbol{x}$ is isolated in $\mathcal{S}_f$ and the Hes-sian is nonzero at $\boldsymbol{x}$. In Section 11, we summarize the convergence results in the form of a commented Venn diagram and we put them in perspective with those for nonsmooth MM.

The second part (Sections 12–14) is concerned with the implementation of inexact QMM and its numerical behavior. Section 12 describes a numerically stable implementation using truncated CG. The proposed algorithm includes an inner stopping criterion that controls the contraction number γ at the cost of delaying the termination of CG by a small number of iterations, say k. Aside from the outer termination tolerance, γ and k are the only control parameters of the algorithm. In Section 13 we detail the construction of surrogates for multidimensional scal-ing and regularized inversion, and we discuss the convergence properties specific to these problems. Section 14 concludes the paper with numerical experiments on

graph layout (an instance of multidimensional scaling) and X-ray tomography (an instance of regularized inversion). The main takeaways are the robustness to the choice of γ and k and the closeness of the practical solutions produced by inexact QMM to the limit solutions produced by exact QMM.

In the end, the advantages of QMM are versatility coupled with strong convergence guarantees, numerical stability, and robustness to control parameters. It has two limitations. First, its performance rests on properly designing the surrogate, which is problem dependent; this is the inevitable price to pay for versatility. Second, the rate of convergence may be only sublinear; but this is not an issue for problems where it is not necessary to locate a minimizer to very high accuracy, as is usually the case in large-scale optimization.

2. Notation

We denote matrices by bold uppercase roman letters (e.g., $\boldsymbol{A}$), vectors by bold lowercase roman letters (e.g., $\boldsymbol{x}$), sets by calligraphic uppercase letters (e.g., $\mathcal{A}$), and set-valued maps by bold Greek letters (e.g., $\boldsymbol{\Phi}$).

The symbol $\|\cdot\|$ denotes the ℓ_2-norm, and $\|\cdot\|_{\boldsymbol{A}}^2$ is the squared ℓ_2-norm weighted by the matrix $\boldsymbol{A}$, that is, $\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 := \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$.

Let $\mathcal{A} \subset \mathbb{R}^n$. If $\mathcal{A}$ is nonempty, the distance from a point $\boldsymbol{x} \in \mathbb{R}^n$ to $\mathcal{A}$ is

$$\operatorname{dist}(\boldsymbol{x}, \mathcal{A}) := \inf_{\boldsymbol{y} \in \mathcal{A}} \|\boldsymbol{y} - \boldsymbol{x}\|. \tag{2.1}$$

A sequence $\{\boldsymbol{x}_p\}$ in $\mathbb{R}^n$ is said to converge to $\mathcal{A}$ if $\lim_p \operatorname{dist}(\boldsymbol{x}_p, \mathcal{A}) = 0$. When $\{\boldsymbol{x}_p\}$ converges in norm to some point, we simply say that it converges. The limit set of $\{\boldsymbol{x}_p\}$ is denoted by $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ (that is, $\boldsymbol{x} \in \mathcal{L}_{\{\boldsymbol{x}_p\}}$ if and only if $\{\boldsymbol{x}_p\}$ has a subsequence converging to $\boldsymbol{x}$).

The interior of $\mathcal{A}$ is denoted by $\mathcal{A}^\circ$ and its boundary by $\partial\mathcal{A}$ (so $\partial\mathcal{A} = \overline{\mathcal{A}} \setminus \mathcal{A}^\circ$, where $\overline{\mathcal{A}}$ is the closure of $\mathcal{A}$). We let $\mathscr{B}(\boldsymbol{x}, r)$ and $\overline{\mathscr{B}}(\boldsymbol{x}, r)$ denote, respectively, the open and closed ℓ_2-balls with center $\boldsymbol{x}$ and radius r.

Let f be a real-valued function on $\mathbb{R}^n$. Given a binary relation $\mathcal{R}$ on $\mathbb{R}$ and a real number α, we use the shortcut notation

$$\{f \mathcal{R} \alpha\} := \big\{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) \mathcal{R} \alpha\big\}. \tag{2.2}$$

For example, $\{f = \alpha\}$ and $\{f \leqslant \alpha\}$ are the level and sublevel sets of f at height α, respectively. Similarly, the inverse image of a set $\mathcal{I} \subset \mathbb{R}$ is denoted by $\{f \in \mathcal{I}\}$.

If f is differentiable, we denote its set of stationary points by $\mathcal{S}_f$ (that is, $\boldsymbol{x} \in \mathcal{S}_f$ if and only if $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$, where ∇f is the gradient of f). Finally, $\mathscr{C}^1$ and $\mathscr{C}^2$ are, respectively, the classes of continuously and twice continuously differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$.

3. Quadratic majorization-minimization algorithms

Definition 3.1. A (quadratic) *surrogate* for an objective $f : \mathbb{R}^n \to \mathbb{R}$ is a family of functions $f(\cdot|\cdot) := \{f(\cdot|\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}\}_{\boldsymbol{x} \in \mathbb{R}^n}$ satisfying the following conditions for all $\boldsymbol{x}$:

(i) $f(\boldsymbol{x}|\boldsymbol{x}) = f(\boldsymbol{x})$ (tangency).

(ii) $f(\boldsymbol{y}|\boldsymbol{x}) \geqslant f(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathbb{R}^n$ (domination).

(iii) $f(\cdot|\boldsymbol{x})$ is a positive definite quadratic function.

A function $f(\cdot|\boldsymbol{x})$ having these properties is called a *surrogate function*.

In other words, there are functions $\boldsymbol{w} : \mathbb{R}^n \to \mathbb{R}^n$ and $\boldsymbol{W} : \mathbb{R}^n \to \mathrm{Sym}^+(n)$ (the set of $n \times n$ symmetric positive definite matrices) such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\begin{aligned} f(\boldsymbol{y}\,|\,\boldsymbol{x}) \;&=\; f(\boldsymbol{x}) + (\boldsymbol{y}-\boldsymbol{x})^T \boldsymbol{w}(\boldsymbol{x}) + \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2_{\boldsymbol{W}(\boldsymbol{x})} \\ &\geqslant\; f(\boldsymbol{y}). \end{aligned} \tag{3.1}$$

We call the functions $\boldsymbol{w}$ and $\boldsymbol{W}$ the (first- and second-order, respectively) *weighting functions* of the surrogate.

The following proposition shows that when f is differentiable the affine component of the surrogate function is the first-order Taylor approximation of f.

Proposition 3.2. *Let $\boldsymbol{x} \in \mathbb{R}^n$. If f is differentiable at $\boldsymbol{x}$ then $\boldsymbol{w}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$.*

Proof. We have for all $\boldsymbol{h} \in \mathbb{R}^n$ that

$$f(\boldsymbol{x}+\boldsymbol{h}) \;=\; f(\boldsymbol{x}) + \boldsymbol{h}^T \nabla f(\boldsymbol{x}) + \underset{\boldsymbol{h}\to\boldsymbol{0}}{o}(\|\boldsymbol{h}\|).$$

Furthermore, by the differentiability of $f(\cdot\,|\,\boldsymbol{x})$ at $\boldsymbol{x}$,

$$f(\boldsymbol{x}+\boldsymbol{h}\,|\,\boldsymbol{x}) \;=\; f(\boldsymbol{x}\,|\,\boldsymbol{x}) + \boldsymbol{h}^T \boldsymbol{w}(\boldsymbol{x}) + \underset{\boldsymbol{h}\to\boldsymbol{0}}{o}(\|\boldsymbol{h}\|).$$

Using the tangency and domination properties, it follows that

$$0 \;\leqslant\; f(\boldsymbol{x}+\boldsymbol{h}\,|\,\boldsymbol{x}) - f(\boldsymbol{x}+\boldsymbol{h}) \;=\; \boldsymbol{h}^T\big(\boldsymbol{w}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\big) + \underset{\boldsymbol{h}\to\boldsymbol{0}}{o}(\|\boldsymbol{h}\|).$$

Hence, for all $\boldsymbol{h} \neq \boldsymbol{0}$,

$$0 \;\leqslant\; (\boldsymbol{h}^T/\|\boldsymbol{h}\|)\big(\boldsymbol{w}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\big) + \varepsilon(\boldsymbol{h}),$$

where $\varepsilon(\boldsymbol{h})$ goes to zero as $\boldsymbol{h} \to \boldsymbol{0}$. Consider the sequence $(\boldsymbol{h}_p)_p$ defined by $\boldsymbol{h}_p := -(1/p)\big(\boldsymbol{w}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\big)$ for all $p \geqslant 1$. Then

$$0 \;\leqslant\; \lim_p \big(-\|\boldsymbol{w}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\| + \varepsilon(\boldsymbol{h}_p)\big) = -\|\boldsymbol{w}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|,$$

so $\boldsymbol{w}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})$ must be zero. □

For any $\gamma \in [0,1)$, we define the set-valued map $\boldsymbol{\Phi}_\gamma : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ by

$$\boldsymbol{\Phi}_\gamma(\boldsymbol{x}) \;:=\; \big\{\, \boldsymbol{y} \in \mathbb{R}^n :\; f(\boldsymbol{y}\,|\,\boldsymbol{x}) - \min f(\cdot\,|\,\boldsymbol{x}) \leqslant \gamma d(\boldsymbol{x}) \,\big\}, \tag{3.2a}$$

$$d(\boldsymbol{x}) \;:=\; f(\boldsymbol{x}) - \min f(\cdot\,|\,\boldsymbol{x}). \tag{3.2b}$$

Definition 3.3. A sequence $\{\boldsymbol{x}_p\}_{p\in\mathbb{N}} \subset \mathbb{R}^n$ is called a *QMM sequence* if there is a constant $\gamma \in [0,1)$ (called the *contraction number)* such that

$$\boldsymbol{x}_{p+1} \;\in\; \boldsymbol{\Phi}_\gamma(\boldsymbol{x}_p) \quad \text{for all } p \in \mathbb{N}. \tag{3.3}$$

If, in addition, the weighting functions $\boldsymbol{w}$ and $\boldsymbol{W}$ are continuous, we say that $\{\boldsymbol{x}_p\}$ is a *$\mathscr{C}^0$-QMM sequence.* The iteration (3.3) starting from a given point $\boldsymbol{x}_0$ is called a *($\mathscr{C}^0$-)QMM algorithm.*

The gradient of the surrogate function is given by

$$\nabla f(\cdot\,|\,\boldsymbol{x})(\boldsymbol{y}) \;=\; \boldsymbol{w}(\boldsymbol{x}) + \boldsymbol{W}(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}). \tag{3.4}$$

Therefore, in the special case $\gamma = 0$, we have

$$\boldsymbol{\Phi}_0(\boldsymbol{x}) \;=\; \operatorname*{arg\,min}_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{y}\,|\,\boldsymbol{x}) \;=\; \{\,\phi(\boldsymbol{x})\,\}, \tag{3.5a}$$

$$\phi(\boldsymbol{x}) \;:=\; \boldsymbol{x} - \boldsymbol{W}(\boldsymbol{x})^{-1}\boldsymbol{w}(\boldsymbol{x}). \tag{3.5b}$$

The QMM process defined by the single-valued map $\boldsymbol{\Phi}_0$ is called *exact*, as opposed to the general *inexact* process in which the contraction number sets the accuracy of the approximation to $\phi(\boldsymbol{x})$.

4. Elementary properties

In this section we give some terminology and some miscellaneous results that will be needed later.

Definition 4.1. Let $\boldsymbol{\Phi}$ be a set-valued map from $\mathbb{R}^n$ to the subsets of $\mathbb{R}^n$.

(i) A point $\boldsymbol{x} \in \mathbb{R}^n$ is called a *fixed point* of $\boldsymbol{\Phi}$ if $\boldsymbol{\Phi}(\boldsymbol{x}) = \{\boldsymbol{x}\}$. The set of fixed points of $\boldsymbol{\Phi}$ is denoted by $\mathcal{F}_{\boldsymbol{\Phi}}$.

(ii) A set $\mathcal{A} \subset \mathbb{R}^n$ is said to be *stable* with respect to $\boldsymbol{\Phi}$ if

$$\boldsymbol{\Phi}(\mathcal{A}) := \bigcup_{x \in \mathcal{A}} \boldsymbol{\Phi}(\boldsymbol{x}) \subset \mathcal{A}.$$

(iii) The map $\boldsymbol{\Phi}$ is said to be *strictly monotonic* with respect to f if for all $\boldsymbol{x} \in \mathbb{R}^n \setminus \mathcal{F}_{\boldsymbol{\Phi}}$, we have $\boldsymbol{\Phi}(\boldsymbol{x}) \subset \{ f < f(\boldsymbol{x}) \}$.

(iv) The map $\boldsymbol{\Phi}$ is said to be *outer semicontinuous* if its graph, that is, the set

$$\operatorname{graph} \boldsymbol{\Phi} := \big\{ (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^n : \boldsymbol{y} \in \boldsymbol{\Phi}(\boldsymbol{x}) \big\},$$

is closed.

Proposition 4.2. *For any $\gamma \in [0,1)$, $\boldsymbol{x} \in \mathcal{F}_{\boldsymbol{\Phi}_\gamma}$ if and only if $\boldsymbol{w}(\boldsymbol{x}) = \boldsymbol{0}$.*

Proof. We have

$$d(\boldsymbol{x}) = 0 \iff \operatorname*{arg\,min}_{\boldsymbol{y} \in \mathbb{R}^n} f(\boldsymbol{y}|\boldsymbol{x}) = \{\boldsymbol{x}\} \iff \boldsymbol{w}(\boldsymbol{x}) = \boldsymbol{0} \tag{4.1}$$

(the first equivalence follows from the fact that $f(\cdot|\boldsymbol{x})$ is a positive definite quadratic function such that $f(\boldsymbol{x}|\boldsymbol{x}) = f(\boldsymbol{x})$ and the second equivalence follows from (3.5)). Let $\alpha(\boldsymbol{x}) := \gamma d(\boldsymbol{x}) + \min f(\cdot|\boldsymbol{x})$. Then

$$\begin{aligned} \boldsymbol{x} \in \mathcal{F}_{\boldsymbol{\Phi}_\gamma} &\iff \big\{ f(\cdot|\boldsymbol{x}) \leqslant \alpha(\boldsymbol{x}) \big\} = \{\boldsymbol{x}\} \\ &\iff \alpha(\boldsymbol{x}) = \min f(\cdot|\boldsymbol{x}) \ \text{ and } \ \operatorname*{arg\,min}_{\boldsymbol{y} \in \mathbb{R}^n} f(\boldsymbol{y}|\boldsymbol{x}) = \{\boldsymbol{x}\}. \end{aligned}$$

Since $d(\boldsymbol{x}) = 0$ implies that $\alpha(\boldsymbol{x}) = \min f(\cdot|\boldsymbol{x})$, it follows from (4.1) that $\boldsymbol{x} \in \mathcal{F}_{\boldsymbol{\Phi}_\gamma}$ is equivalent to $\boldsymbol{w}(\boldsymbol{x}) = \boldsymbol{0}$. □

Propositions 3.2 and 4.2 yield the following corollary.

Corollary 4.3. *If f is differentiable then $\mathcal{F}_{\boldsymbol{\Phi}_\gamma} = \mathcal{S}_f$ for all $\gamma \in [0,1)$.*

Recall that a connected component $\mathcal{C}$ of a set $\mathcal{A} \subset \mathbb{R}^n$ is a maximal connected subset of $\mathcal{A}$ (that is, $\mathcal{C}$ is connected and $\mathcal{C}$ is not a proper subset of a connected subset of $\mathcal{A}$).

Definition 4.4. A connected component of a sublevel set of f is called a *basin*.

Proposition 4.5. *The basins of f are stable with respect to $\boldsymbol{\Phi}_\gamma$ for all $\gamma \in [0,1)$.*

Proof. Let $\alpha \in \mathbb{R}$ be such that $\{ f \leqslant \alpha \}$ is nonempty and let $\mathcal{C}$ be a connected component of $\{ f \leqslant \alpha \}$. Let $\boldsymbol{x} \in \mathcal{C}$ and let $\boldsymbol{y} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x})$. Since $f(\cdot|\boldsymbol{x})$ is a positive definite quadratic function, the set

$$\boldsymbol{\Phi}_1(\boldsymbol{x}) = \big\{ \boldsymbol{y} \in \mathbb{R}^n : f(\boldsymbol{y}|\boldsymbol{x}) \leqslant f(\boldsymbol{x}) \big\}$$

is an ellipsoid and is therefore convex. Since $\boldsymbol{x} \in \boldsymbol{\Phi}_1(\boldsymbol{x})$ and $\boldsymbol{\Phi}_\gamma(\boldsymbol{x}) \subset \boldsymbol{\Phi}_1(\boldsymbol{x}) \subset \{\, f \leqslant \alpha \,\}$, it follows that

$$[\boldsymbol{x}, \boldsymbol{y}] \;:=\; \big\{\, \boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}) : t \in [0,1] \,\big\} \;\subset\; \{\, f \leqslant \alpha \,\}.$$

Consequently, $\mathcal{C} \cup [\boldsymbol{x}, \boldsymbol{y}]$ is a connected subset of $\{\, f \leqslant \alpha \,\}$, so $\boldsymbol{y}$ must be in $\mathcal{C}$. □

Proposition 4.6. *For any $\gamma \in [0,1)$, $\boldsymbol{\Phi}_\gamma$ is strictly monotonic with respect to f.*

Proof. For any $\boldsymbol{x} \in \mathbb{R}^n$, we have

$$\min f(\cdot\,|\,\boldsymbol{x}) \;=\; f\big(\phi(\boldsymbol{x})\,\big|\,\boldsymbol{x}\big) \;=\; f(\boldsymbol{x}) - \tfrac{1}{2}\|\boldsymbol{w}(\boldsymbol{x})\|^2_{\boldsymbol{W}(\boldsymbol{x})^{-1}}. \tag{4.2}$$

Suppose $\boldsymbol{x} \notin \mathcal{F}_{\boldsymbol{\Phi}_\gamma}$. By Proposition 4.2 we have $\boldsymbol{w}(\boldsymbol{x}) \neq \boldsymbol{0}$ and so $\min f(\cdot\,|\,\boldsymbol{x}) < f(\boldsymbol{x})$. Thus, for any $\boldsymbol{y} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x})$ we have

$$\begin{aligned} f(\boldsymbol{y}) \;\leqslant\; f(\boldsymbol{y}\,|\,\boldsymbol{x}) \;&\leqslant\; \gamma d(\boldsymbol{x}) + \min f(\cdot\,|\,\boldsymbol{x}) \\ &=\; \gamma f(\boldsymbol{x}) + (1-\gamma)\min f(\cdot\,|\,\boldsymbol{x}) \\ &<\; f(\boldsymbol{x}). \end{aligned}$$

□

Corollary 4.7. *Let $\{\boldsymbol{x}_p\}$ be a QMM sequence. Then $\{f(\boldsymbol{x}_p)\}$ is decreasing, and it is strictly decreasing as long as $\boldsymbol{x}_p \notin \mathcal{F}_{\boldsymbol{\Phi}_\gamma}$.*

Definition 4.8. Let $\mathcal{A}$ be a nonempty subset of $\mathbb{R}^n$.

(i) $\mathcal{A}$ is said to be *flat* with respect to f (or just *flat*) if f is constant on $\mathcal{A}$.

(ii) $\mathcal{A}$ is called a *continuum* if it is compact and connected.

(In particular, a singleton is a flat continuum.)

Proposition 4.9. *Let $\{\boldsymbol{x}_p\}$ be a QMM sequence with continuous objective. If $\{\boldsymbol{x}_p\}$ is bounded and $\{\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\}$ converges to $\boldsymbol{0}$, then the limit set $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is a flat continuum.*

Proof. Since $\{\boldsymbol{x}_p\}$ is bounded and the set of limit points of a sequence is closed, $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is nonempty and compact. Since $\boldsymbol{x}_{p+1} - \boldsymbol{x}_p$ goes to zero, it follows from [17, Theorem 26.1] that $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is connected.

Suppose $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is not flat, that is, there exist $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{L}_{\{\boldsymbol{x}_p\}}$ such that $f(\boldsymbol{x}) < f(\boldsymbol{y})$. Let $\rho := \frac{1}{2}(f(\boldsymbol{y}) - f(\boldsymbol{x}))$. Since f is continuous, there exists $r > 0$ such that $f(\mathscr{B}(\boldsymbol{x}, r)) \subset \mathscr{B}(f(\boldsymbol{x}), \rho)$ and $f(\mathscr{B}(\boldsymbol{y}, r)) \subset \mathscr{B}(f(\boldsymbol{y}), \rho)$. Furthermore, there exists integers p and $q > p$ such that $\boldsymbol{x}_p \in \mathscr{B}(\boldsymbol{x}, r)$ and $\boldsymbol{x}_q \in \mathscr{B}(\boldsymbol{y}, r)$. Therefore

$$\begin{aligned} f(\boldsymbol{x}_q) - f(\boldsymbol{x}_p) \;&=\; f(\boldsymbol{y}) - f(\boldsymbol{x}) + f(\boldsymbol{x}_q) - f(\boldsymbol{y}) - f(\boldsymbol{x}_p) + f(\boldsymbol{x}) \\ &\geqslant\; 2\rho - |f(\boldsymbol{x}_q) - f(\boldsymbol{y})| - |f(\boldsymbol{x}_p) - f(\boldsymbol{x})| \\ &>\; 0, \end{aligned}$$

which contradicts Corollary 4.7. So $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ must be flat. □

Proposition 4.10. *If the weighting functions $\boldsymbol{w}$ and $\boldsymbol{W}$ are continuous, then for any $\gamma \in [0,1)$, $\boldsymbol{\Phi}_\gamma$ is outer semicontinuous.*

Proof. If $\boldsymbol{w}$ and $\boldsymbol{W}$ are continuous, then the function $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^n \longmapsto f(\boldsymbol{y}\,|\,\boldsymbol{x})$ is continuous. Furthermore, since $\boldsymbol{W}(\boldsymbol{x})$ is invertible for all $\boldsymbol{x}$, the function $\boldsymbol{x} \longmapsto \boldsymbol{W}(\boldsymbol{x})^{-1}$ is continuous and it follows from (4.2) that $\boldsymbol{x} \longmapsto \gamma d(\boldsymbol{x}) + \min f(\cdot\,|\,\boldsymbol{x})$ is continuous.

Let $\{(\boldsymbol{x}_p, \boldsymbol{y}_p)\}$ be a sequence in graph $\boldsymbol{\Phi}_\gamma$ and suppose that $\{\boldsymbol{x}_p\}$ and $\{\boldsymbol{y}_p\}$ converge to some points $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. We have

$$f(\boldsymbol{y}\,|\,\boldsymbol{x}) \;=\; \lim_p f(\boldsymbol{y}_p\,|\,\boldsymbol{x}_p) \;\leqslant\; \lim_p \big(\gamma d(\boldsymbol{x}_p) + \min f(\cdot\,|\,\boldsymbol{x}_p)\big) \;=\; \gamma d(\boldsymbol{x}) + \min f(\cdot\,|\,\boldsymbol{x}),$$

so $(\boldsymbol{x}, \boldsymbol{y}) \in \operatorname{graph} \boldsymbol{\Phi}_\gamma$. □

5. Subconvergence

From now on, we focus on bounded $\mathscr{C}^0$-QMM sequences with differentiable objectives. Theorem 5.2 below shows that the limit points of such a sequence are stationary points of the objective and form a flat continuum; it will be used in the proofs of the convergence theorems in Sections 6 through 10. (Note that by Proposition 4.5, a QMM sequence is bounded whenever its starting point is in a bounded basin. In particular, QMM sequences with coercive objectives are bounded.)

Lemma 5.1. *Let $\{\boldsymbol{x}_p\}$ be a bounded sequence in $\mathbb{R}^n$.*

(i) *For any $\boldsymbol{x} \in \mathcal{L}_{\{\boldsymbol{x}_p\}}$, $\{\boldsymbol{x}_p\}$ has a subsequence $\{\boldsymbol{x}_{p_k}\}$ converging to $\boldsymbol{x}$ such that $\{\boldsymbol{x}_{p_k+1}\}$ is convergent.*

(ii) *If $\{\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\}$ does not converge to $\mathbf{0}$, then $\{\boldsymbol{x}_p\}$ has a convergent subsequence $\{\boldsymbol{x}_{p_k}\}$ such that $\{\boldsymbol{x}_{p_k+1}\}$ is convergent and $\lim_k \|\boldsymbol{x}_{p_k+1} - \boldsymbol{x}_{p_k}\| > 0$.*

Proof. Since $\boldsymbol{x}$ is a limit point, $\{\boldsymbol{x}_p\}$ has a subsequence $\{\boldsymbol{x}_{q_i}\}$ converging to $\boldsymbol{x}$. Because $\{\boldsymbol{x}_p\}$ is bounded, $\{\boldsymbol{x}_{q_i+1}\}$ has a convergent subsequence $\{\boldsymbol{x}_{q_{i_k}+1}\}$. Letting $p_k := q_{i_k}$ proves the first assertion.

Since $\|\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\|$ does not tend to zero, there exist $\varepsilon > 0$ and a subsequence $\{\boldsymbol{x}_{q_i}\}$ such that $\|\boldsymbol{x}_{q_i+1} - \boldsymbol{x}_{q_i}\| > \varepsilon$ for all i. Because $\{\boldsymbol{x}_p\}$ is bounded, $\{\boldsymbol{x}_{q_i}\}$ has a convergent subsequence $\{\boldsymbol{x}_{q_{i_j}}\}$, and in turn $\{\boldsymbol{x}_{q_{i_j}+1}\}$ has a convergent subsequence indexed by $q_{i_{j_k}} + 1$. Letting $p_k := q_{i_{j_k}}$ proves the second assertion. □

Theorem 5.2. *Let $\{\boldsymbol{x}_p\}$ be a $\mathscr{C}^0$-QMM sequence with differentiable objective f. If $\{\boldsymbol{x}_p\}$ is bounded then*

(i) *$\emptyset \neq \mathcal{L}_{\{\boldsymbol{x}_p\}} \subset \mathcal{S}_f$;*

(ii) *$\{\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\}$ converges to $\mathbf{0}$, so $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is a flat continuum;*

(iii) *$\{f(\boldsymbol{x}_p)\}$ decreases to the value of f on $\mathcal{L}_{\{\boldsymbol{x}_p\}}$, and it is strictly decreasing as long as $\boldsymbol{x}_p \notin \mathcal{S}_f$.*

Proof. First note that $\mathcal{F}_{\boldsymbol{\Phi}_\gamma} = \mathcal{S}_f$ (Corollary 4.3) and that $\boldsymbol{\Phi}_\gamma$ is strictly monotonic and outer semicontinuous (Propositions 4.6 and 4.10).

(i) Since $\{\boldsymbol{x}_p\}$ is bounded, it has at least one limit point. Let $\boldsymbol{x} \in \mathcal{L}_{\{\boldsymbol{x}_p\}}$ and suppose $\boldsymbol{x} \notin \mathcal{F}_{\boldsymbol{\Phi}_\gamma}$. From Lemma 5.1(i) there is a subsequence $\{\boldsymbol{x}_{p_k}\}$ converging to $\boldsymbol{x}$ such that $\{\boldsymbol{x}_{p_k+1}\}$ converges to some point $\boldsymbol{y}$. Since $\boldsymbol{x}_{p_k+1} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x}_{p_k})$ for all k, it follows from outer semicontinuity that $\boldsymbol{y} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x})$. Hence $f(\boldsymbol{y}) < f(\boldsymbol{x})$ by strict monotonicity. But $\{f(\boldsymbol{x}_p)\}$ is decreasing, so

$$f(\boldsymbol{x}) \;=\; \lim_k f(\boldsymbol{x}_{p_{k+1}}) \;\leqslant\; \lim_k f(\boldsymbol{x}_{p_k+1}) \;=\; f(\boldsymbol{y}),$$

which is a contradiction. Thus $\boldsymbol{x} \in \mathcal{F}_{\boldsymbol{\Phi}_\gamma}$.

(ii) Suppose that $\{\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\}$ does not converge to $\mathbf{0}$. From Lemma 5.1(ii) there is a convergent subsequence $\{\boldsymbol{x}_{p_k}\}$ such that $\{\boldsymbol{x}_{p_k+1}\}$ is also convergent and

$$\boldsymbol{x} \;:=\; \lim_k \boldsymbol{x}_{p_k} \;\neq\; \lim_k \boldsymbol{x}_{p_k+1} \;=:\; \boldsymbol{y}.$$

But $\boldsymbol{x} \in \mathcal{F}_{\boldsymbol{\Phi}_\gamma}$ by (i) and $\boldsymbol{y} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x})$ by outer semicontinuity; so $\boldsymbol{x} = \boldsymbol{y}$, which is a contradiction. Therefore $\boldsymbol{x}_{p+1} - \boldsymbol{x}_p$ goes to zero and it follows from Proposition 4.9 that $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is a flat continuum.

(iii) This result follows immediately from Corollary 4.7 and (ii). □

6. Global convergence

As a consequence of subconvergence, the next theorem shows that a $\mathscr{C}^0$-QMM sequence $\{\boldsymbol{x}_p\}$ converges when the stationary points of the objective f are isolated in its level sets. If not, we are guaranteed that $\{\boldsymbol{x}_p\}$ converges to $\mathcal{S}_f$, which has two implications: the gradient norm goes to zero when f is $\mathscr{C}^1$, and $\{\boldsymbol{x}_p\}$ converges to the set of global minimizers when f is convex and coercive.

Definition 6.1. Let $\mathcal{A}$ be a nonempty subset of $\mathbb{R}^n$. We say that $\mathcal{A}$ is *level-discrete* (with respect to f) if for any $\alpha \in \mathbb{R}$, $\mathcal{A} \cap \{f = \alpha\}$ is discrete or empty.

Theorem 6.2. *Let $\{\boldsymbol{x}_p\}$ be a $\mathscr{C}^0$-QMM sequence with differentiable objective f and suppose that $\{\boldsymbol{x}_p\}$ is bounded.*

(i) *If $\mathcal{S}_f$ is level-discrete, then $\{\boldsymbol{x}_p\}$ converges to a stationary point of f.*

(ii) *$\{\boldsymbol{x}_p\}$ converges to $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ and hence to $\mathcal{S}_f$.*

Proof. First note that $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is a flat continuum contained in $\mathcal{S}_f$, by Theorem 5.2.

(i) Suppose $\mathcal{S}_f$ is level-discrete. Then $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is discrete (as a flat subset of $\mathcal{S}_f$) and hence is a singleton (otherwise $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is not a continuum). Let $\boldsymbol{x}$ be the unique limit point of $\{\boldsymbol{x}_p\}$ and suppose that $\{\boldsymbol{x}_p\}$ diverges. Then there exist $\varepsilon > 0$ and a subsequence $\{\boldsymbol{x}_{p_k}\}$ such that $\|\boldsymbol{x}_{p_k} - \boldsymbol{x}\| > \varepsilon$ for all k. Since $\{\boldsymbol{x}_p\}$ is bounded, $\{\boldsymbol{x}_{p_k}\}$ has a further subsequence converging to some point $\boldsymbol{y} \neq \boldsymbol{x}$, contradicting the fact that $\mathcal{L}_{\{\boldsymbol{x}_p\}} = \{\boldsymbol{x}\}$. Thus $\{\boldsymbol{x}_p\}$ converges.

(ii) Suppose $\{\boldsymbol{x}_p\}$ does not converge to $\mathcal{L}_{\{\boldsymbol{x}_p\}}$. Then there exist $\varepsilon > 0$ and a subsequence $\{\boldsymbol{x}_{p_k}\}$ such that $\operatorname{dist}(\boldsymbol{x}_{p_k}, \mathcal{L}_{\{\boldsymbol{x}_p\}}) > \varepsilon$ for all k. Since $\mathcal{L}_{\{\boldsymbol{x}_{p_k}\}} \subset \mathcal{L}_{\{\boldsymbol{x}_p\}}$, it follows that $\mathcal{L}_{\{\boldsymbol{x}_{p_k}\}}$ is empty, which contradicts the boundedness of $\{\boldsymbol{x}_p\}$. Therefore $\{\boldsymbol{x}_p\}$ converges to $\mathcal{L}_{\{\boldsymbol{x}_p\}}$. □

Corollary 6.3. *Let $\{\boldsymbol{x}_p\}$ be a $\mathscr{C}^0$-QMM sequence with differentiable objective f.*

(i) *If $\{\boldsymbol{x}_p\}$ is bounded and f is $\mathscr{C}^1$ then $\{\nabla f(\boldsymbol{x}_p)\}$ converges to $\mathbf{0}$.*

(ii) *If f is convex and coercive, then $\{\boldsymbol{x}_p\}$ converges to $\arg\min f$.*

Proof. (i) It follows from Theorem 6.2(ii) that there is a sequence $\{\boldsymbol{y}_p\} \subset \mathcal{S}_f$ such that $\lim_p \|\boldsymbol{x}_p - \boldsymbol{y}_p\| = 0$. If f is $\mathscr{C}^1$ then ∇f is uniformly continuous on any compact set containing $\{\boldsymbol{x}_p\}$, and hence $\lim_p \|\nabla f(\boldsymbol{x}_p)\| = \lim_p \|\nabla f(\boldsymbol{x}_p) - \nabla f(\boldsymbol{y}_p)\| = 0$.

(ii) If f is coercive then its sublevel sets are bounded, and thus so are its basins. By Proposition 4.5, $\{\boldsymbol{x}_p\}$ is bounded and it follows from Theorem 6.2(ii) that $\{\boldsymbol{x}_p\}$ converges to $\mathcal{S}_f$. If, in addition, f is convex, then $\mathcal{S}_f$ is the set of global minimizers of f (see, e.g., [18, Theorem 7.4-4]). □

7. Local convergence

Theorem 6.2 does not exclude the possibility that $\mathscr{C}^0$-QMM sequences be attracted by saddle points. Nevertheless, Theorem 7.3 below shows that stationary points isolated in bounded basins are strict local minimizers and are points of attraction.

Lemma 7.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and let $\mathcal{B}$ be a basin of f. Then there is an $\alpha \in \mathbb{R}$ such that $\mathcal{B} \subset \{ f \leqslant \alpha \}$ and $\partial\mathcal{B} \subset \{ f = \alpha \}$.*

Proof. By definition, $\mathcal{B}$ is a connected component of a sublevel set $\{ f \leqslant \alpha \}$. Suppose there is a point $\boldsymbol{x} \in \partial\mathcal{B} \cap \{ f < \alpha \}$. Since f is continuous, there is an $r > 0$ such that $\mathscr{B}(\boldsymbol{x}, r) \subset \{ f < \alpha \}$, and hence $\mathcal{B} \cup \overline{\mathscr{B}}(\boldsymbol{x}, r)$ is a connected subset of $\{ f \leqslant \alpha \}$. Since $\mathscr{B}(\boldsymbol{x}, r) \not\subset \mathcal{B}$ (for otherwise $\boldsymbol{x} \in \mathcal{B}^\circ$), we have a contradiction. □

Lemma 7.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and let $\mathcal{B}$ be a bounded basin of f such that $\mathcal{B}^\circ \cap \mathcal{S}_f$ is a singleton $\{\boldsymbol{x}\}$. Then* $\arg\min_{\mathcal{B}} f = \{\boldsymbol{x}\}$.

Proof. Let α be as in Lemma 7.1 and suppose $\min_{\mathcal{B}} f = \alpha$. Then $\mathcal{B}$ is flat and so $\mathcal{B}^\circ \cap \mathcal{S}_f = \mathcal{B}^\circ$. This contradicts the fact that $\mathcal{B}^\circ \cap \mathcal{S}_f$ is a singleton. Thus $\arg\min_{\mathcal{B}} f = \arg\min_{\mathcal{B}^\circ} f$. Since $\mathcal{B}$ is compact and $\arg\min_{\mathcal{B}^\circ} f \subset \mathcal{B}^\circ \cap \mathcal{S}_f = \{\boldsymbol{x}\}$, we conclude that $\arg\min_{\mathcal{B}} f = \{\boldsymbol{x}\}$. □

Theorem 7.3. *Let $\{\boldsymbol{x}_p\}$ be a $\mathscr{C}^0$-QMM sequence with differentiable objective f and let $\mathcal{B}$ be a bounded basin of f such that $\mathcal{B}^\circ \cap \mathcal{S}_f = \{\boldsymbol{x}\}$. The point $\boldsymbol{x}$ is a strict local minimizer, and if the starting point $\boldsymbol{x}_0$ is in $\mathcal{B}^\circ$ then $\{\boldsymbol{x}_p\}$ converges to $\boldsymbol{x}$.*

Proof. It follows immediately from Lemma 7.2 that $\boldsymbol{x}$ is a strict local minimizer. Suppose that $\boldsymbol{x}_0 \in \mathcal{B}^\circ$. Then $\{\boldsymbol{x}_p\} \subset \mathcal{B}$ (Proposition 4.5) and hence converges to $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ (Theorem 6.2(ii)), so we must show that $\boldsymbol{x}$ is the only limit point of $\{\boldsymbol{x}_p\}$.

Using Theorem 5.2, we have that $\emptyset \neq \mathcal{L}_{\{\boldsymbol{x}_p\}} \subset \mathcal{B} \cap \mathcal{S}_f \subset \partial\mathcal{B} \cup \{\boldsymbol{x}\}$. Moreover, $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is flat and we have from Lemmas 7.1 and 7.2 that $f(\boldsymbol{x}) < \alpha$, where α is the value of f on $\partial\mathcal{B}$. So either $\mathcal{L}_{\{\boldsymbol{x}_p\}} \subset \partial\mathcal{B}$ or $\mathcal{L}_{\{\boldsymbol{x}_p\}} = \{\boldsymbol{x}\}$.

If $\boldsymbol{x}_0 \in \mathcal{S}_f$ then $\boldsymbol{x}_0 = \boldsymbol{x}$, which implies that $\boldsymbol{x}_p = \boldsymbol{x}$ for all p (Corollary 4.3) and hence $\mathcal{L}_{\{\boldsymbol{x}_p\}} = \{\boldsymbol{x}\}$. If $\boldsymbol{x}_0 \notin \mathcal{S}_f$ then $f(\boldsymbol{x}_0) < \alpha$ (for otherwise $\boldsymbol{x}_0 \in \mathcal{B}^\circ \cap \arg\max_{\mathcal{B}} f \subset \mathcal{S}_f$), and it follows from Theorem 5.2(iii) that the value of f on $\mathcal{L}_{\{\boldsymbol{x}_p\}}$ is smaller than α; therefore $\mathcal{L}_{\{\boldsymbol{x}_p\}} \not\subset \partial\mathcal{B}$, which completes the proof. □

8. Finite-length convergence

In this section we extend the previous convergence properties by showing that $\mathscr{C}^0$-QMM algorithms generate finite-length trajectories when the objective satisfies the KL inequality. We also discuss the scope of this assumption.

For any $\rho \in \mathbb{R}$, we denote by $\mathscr{D}(\rho)$ the class of continuously differentiable functions $\psi : (0, \rho) \to (0, +\infty)$ that are concave and strictly increasing.

Definition 8.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, let $\boldsymbol{x} \in \mathbb{R}^n$, and set $f_{\boldsymbol{x}} := f - f(\boldsymbol{x})$. We say that f has the *KL property* at $\boldsymbol{x}$ (or that f satisfies the *KL inequality* at $\boldsymbol{x}$) if there are positive constants r and ρ and a function $\psi \in \mathscr{D}(\rho)$ such that

$$\|\nabla(\psi \circ f_{\boldsymbol{x}})(\boldsymbol{y})\| \geqslant 1 \quad \text{for all } \boldsymbol{y} \in \mathscr{B}(\boldsymbol{x}, r) \cap \{ f_{\boldsymbol{x}} \in (0, \rho) \}. \tag{8.1}$$

We call f a *KL function* if it has the KL property at all points.

The KL property holds at every point $\boldsymbol{x} \notin \mathcal{S}_f$ and every maximizer. If $\boldsymbol{x}$ is a strict minimizer, the KL property means that f can be concavely distorted into a function that is steep around $\boldsymbol{x}$ (the distortion ψ is called a *desingularizing function*). The following lemma shows that the KL property is uniform on flat compact sets.

Lemma 8.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, let $\mathcal{C}$ be a flat compact set, let α be the value of f on $\mathcal{C}$, and set $f_\alpha := f - \alpha$. Suppose that f has the KL property at all points of $\mathcal{C}$. Then there are positive constants r and ρ and a function $\psi \in \mathscr{D}(\rho)$ such that*

$$\|\nabla(\psi \circ f_\alpha)(\boldsymbol{y})\| \geqslant 1 \quad \text{for all } \boldsymbol{y} \in \mathscr{B}(\mathcal{C}, r) \cap \{ f_\alpha \in (0, \rho) \}, \tag{8.2}$$

where $\mathscr{B}(\mathcal{C}, r) := \{ \operatorname{dist}(\cdot, \mathcal{C}) < r \}$.

Proof. For every $\boldsymbol{x} \in \mathcal{C}$, there are positive constants $r_{\boldsymbol{x}}$ and $\rho_{\boldsymbol{x}}$ and a function $\psi_{\boldsymbol{x}} \in \mathscr{D}(\rho_{\boldsymbol{x}})$ such that

$$\|\nabla(\psi_{\boldsymbol{x}} \circ f_\alpha)(\boldsymbol{y})\| \geqslant 1 \quad \text{for all } \boldsymbol{y} \in \mathscr{B}(\boldsymbol{x}, r_{\boldsymbol{x}}) \cap \{ f_\alpha \in (0, \rho_{\boldsymbol{x}}) \}. \tag{8.3}$$

Since $\mathcal{C}$ is compact, there is a finite set of points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \subset \mathcal{C}$ such that the open balls $\mathscr{B}(\boldsymbol{x}_i, r_{\boldsymbol{x}_i})$, $i = 1, \ldots, k$, cover $\mathcal{C}$. Let $\rho := \min_i \rho_{\boldsymbol{x}_i}$. The function $\varphi : t \in (0, \rho) \longmapsto \max_i \psi'_{\boldsymbol{x}_i}(t)$ is positive, decreasing, and continuous. Moreover, for any $t \in (0, \rho)$ and any $\varepsilon \in (0, t)$, we have

$$\int_\varepsilon^t \varphi \;\leqslant\; \sum_{i=1}^k \int_\varepsilon^t \psi'_{\boldsymbol{x}_i} \;\leqslant\; \sum_{i=1}^k \psi_{\boldsymbol{x}_i}(t),$$

so $\displaystyle\int_0^t \varphi =: \psi(t)$ converges and $\psi \in \mathscr{D}(\rho)$ with $\psi' = \varphi$. Let

$$\mathcal{O} := \bigcup_{i=1}^k \mathscr{B}(\boldsymbol{x}_i, r_{\boldsymbol{x}_i}) \quad \text{and} \quad r := \inf_{\mathbb{R}^n \setminus \mathcal{O}} \operatorname{dist}(\cdot, \mathcal{C}).$$

Suppose that $r = 0$. Then there are sequences $\{\boldsymbol{x}_p\} \subset \mathbb{R}^n \setminus \mathcal{O}$ and $\{\boldsymbol{y}_p\} \subset \mathcal{C}$ such that $\lim_p \|\boldsymbol{x}_p - \boldsymbol{y}_p\| = 0$. The sequence $\{\boldsymbol{y}_p\}$ has a subsequence $\{\boldsymbol{y}_{p_l}\}$ converging to a point $\boldsymbol{y} \in \mathcal{C}$, so $\boldsymbol{x}_{p_l} \in \mathcal{O}$ for sufficiently large l, which is a contradiction. Hence $r > 0$. Let $\boldsymbol{y} \in \mathscr{B}(\mathcal{C}, r)$. There is a point $\boldsymbol{z} \in \mathcal{C}$ such that $\|\boldsymbol{y} - \boldsymbol{z}\| < r$, so $\boldsymbol{y} \in \mathcal{O}$ and thus $\boldsymbol{y} \in \mathscr{B}(\boldsymbol{x}_i, r_{\boldsymbol{x}_i})$ for some i. If $f_\alpha(\boldsymbol{y}) \in (0, \rho)$ then, using (8.3),

$$\begin{aligned} 1 \;\leqslant\; \|\nabla(\psi_{\boldsymbol{x}_i} \circ f_\alpha)(\boldsymbol{y})\| \;&=\; \psi'_{\boldsymbol{x}_i}(f_\alpha(\boldsymbol{y}))\|\nabla f(\boldsymbol{y})\| \\ &\leqslant\; \varphi(f_\alpha(\boldsymbol{y}))\|\nabla f(\boldsymbol{y})\| \;=\; \|\nabla(\psi \circ f_\alpha)(\boldsymbol{y})\|, \end{aligned}$$

which completes the proof. □

We also need the following lemma which gives an upper bound on the difference between two successive iterates of a bounded $\mathscr{C}^0$-QMM sequence.

Lemma 8.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, let $\gamma \in [0, 1)$, and let $\mathcal{C}$ be a compact subset of $\mathbb{R}^n$. If $\boldsymbol{W}$ is continuous, then there is a positive constant δ such that for all $(\boldsymbol{x}, \boldsymbol{y}) \in \operatorname{graph} \boldsymbol{\Phi}_\gamma$ with $\boldsymbol{x} \in \mathcal{C}$, we have*

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \;\geqslant\; \delta\|\nabla f(\boldsymbol{x})\|\,\|\boldsymbol{y} - \boldsymbol{x}\|. \tag{8.4}$$

Proof. Let $\boldsymbol{x} \in \mathcal{C} \setminus \mathcal{S}_f$ and $\boldsymbol{y} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x})$ (if $\boldsymbol{x} \in \mathcal{S}_f$ then (8.4) holds trivially by Corollary 4.3). Substituting (3.1) and (4.2) into the definition of $\boldsymbol{\Phi}_\gamma$ in (3.2), and recalling that $\boldsymbol{w} = \nabla f$ (Proposition 3.2), we obtain

$$\|\boldsymbol{y} - \boldsymbol{x}\|^2_{\boldsymbol{W}(\boldsymbol{x})} + 2(\boldsymbol{y} - \boldsymbol{x})^T \nabla f(\boldsymbol{x}) + (1 - \gamma)\|\nabla f(\boldsymbol{x})\|^2_{\boldsymbol{W}(\boldsymbol{x})^{-1}} \;\leqslant\; 0. \tag{8.5}$$

Hence, using (3.1) again and the domination property,

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \;\geqslant\; \tfrac{1}{2}(1 - \gamma)\|\nabla f(\boldsymbol{x})\|^2_{\boldsymbol{W}(\boldsymbol{x})^{-1}}.$$

For any $\boldsymbol{z} \in \mathbb{R}^n$, let $\lambda_1(\boldsymbol{z})$ and $\lambda_n(\boldsymbol{z})$ denote the smallest and largest eigenvalues of $\boldsymbol{W}(\boldsymbol{z})$, respectively. Since $\mathcal{C}$ is compact and $\boldsymbol{W}$ is a continuous function from $\mathbb{R}^n$ to $\mathrm{Sym}^+(n)$, we have $a := \inf_{\mathcal{C}} \lambda_1 > 0$ and $b := \sup_{\mathcal{C}} \lambda_n < +\infty$, so $\|\boldsymbol{v}\|^2_{\boldsymbol{W}(\boldsymbol{x})} \geqslant a\|\boldsymbol{v}\|^2$ and $\|\boldsymbol{v}\|^2_{\boldsymbol{W}(\boldsymbol{x})^{-1}} \geqslant (1/b)\|\boldsymbol{v}\|^2$ for all $\boldsymbol{v}$. It follows from (8.5) that

$$a\,\frac{\|\boldsymbol{y}-\boldsymbol{x}\|^2}{\|\nabla f(\boldsymbol{x})\|^2} + 2\,\frac{\|\boldsymbol{y}-\boldsymbol{x}\|}{\|\nabla f(\boldsymbol{x})\|} + \frac{1-\gamma}{b} \leqslant 0.$$

The quadratic equation $at^2 - 2t + (1-\gamma)/b = 0$ has positive roots, so there are positive constants c and c' independent of $\boldsymbol{x}$ and $\boldsymbol{y}$ such that

$$c\|\nabla f(\boldsymbol{x})\| \leqslant \|\boldsymbol{y}-\boldsymbol{x}\| \leqslant c'\|\nabla f(\boldsymbol{x})\|. \tag{8.6}$$

Consequently,

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \geqslant \frac{1-\gamma}{2b}\|\nabla f(\boldsymbol{x})\|^2 \geqslant \frac{1-\gamma}{2bc'}\|\nabla f(\boldsymbol{x})\|\,\|\boldsymbol{y}-\boldsymbol{x}\|. \qquad \square$$

The next result shows that a $\mathscr{C}^0$-QMM sequence with a KL objective forms a finite-length trajectory to a stationary point. We denote by $\mathscr{C}_{\mathrm{KL}}$ the class of KL functions on $\mathbb{R}^n$.

Theorem 8.4. *Let $\{\boldsymbol{x}_p\}$ be a bounded $\mathscr{C}^0$-QMM sequence with objective $f \in \mathscr{C}_{\mathrm{KL}}$. Then the series $\sum \|\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\|$ converges and $\lim_p \boldsymbol{x}_p \in \mathcal{S}_f$.*

Proof. If $\boldsymbol{x}_p \in \mathcal{S}_f$ for some p then $\|\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\|$ is eventually zero (Corollary 4.3) and the result is trivial; so we assume that $\{\boldsymbol{x}_p\} \cap \mathcal{S}_f$ is empty. By Theorems 5.2 and 6.2, $\mathcal{L}_{\{\boldsymbol{x}_p\}} =: \mathcal{C}$ is a flat continuum contained in $\mathcal{S}_f$ (so Lemma 8.2 applies), $\{f(\boldsymbol{x}_p)\}$ decreases to the value of f on $\mathcal{C}$ (denoted by α), and $\{\boldsymbol{x}_p\}$ converges to $\mathcal{C}$.

Let $r, \rho \in (0, +\infty)$ and $\psi \in \mathscr{D}(\rho)$ be defined as in Lemma 8.2, and let $t_p := f(\boldsymbol{x}_p) - \alpha$. For sufficiently large p, $\boldsymbol{x}_p \in \mathscr{B}(\mathcal{C}, r)$ and $t_p \in (0, \rho)$, so there is an integer q such that for all $p \geqslant q$, $\psi'(t_p) \geqslant 1/\|\nabla f(\boldsymbol{x}_p)\|$. Hence, since ψ is concave,

$$\psi(t_p) - \psi(t_{p+1}) \geqslant \big(f(\boldsymbol{x}_p) - f(\boldsymbol{x}_{p+1})\big)/\|\nabla f(\boldsymbol{x}_p)\| \quad \text{for all } p \geqslant q.$$

Furthermore, Lemma 8.3 tells us that there exists $\delta > 0$ such that

$$f(\boldsymbol{x}_p) - f(\boldsymbol{x}_{p+1}) \geqslant \delta\|\nabla f(\boldsymbol{x}_p)\|\,\|\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\| \quad \text{for all } p.$$

Therefore, for all $p \geqslant q$,

$$\delta \sum_{k=p}^{\infty} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| \leqslant \sum_{k=p}^{\infty} \big(\psi(t_k) - \psi(t_{k+1})\big) = \psi(t_p) - \lim_{t\to 0}\psi(t) \leqslant \psi(t_p). \tag{8.7}$$

Thus the series $\sum \|\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\|$ converges, so $\{\boldsymbol{x}_p\}$ converges, and since $\mathcal{L}_{\{\boldsymbol{x}_p\}} \subset \mathcal{S}_f$, we have $\lim_p \boldsymbol{x}_p \in \mathcal{S}_f$. $\square$

The KL inequality is named after S. Łojasiewicz, who showed that it holds everywhere for any real analytic functions [19, Proposition 1, p. 92] (see also [20, Proposition 6.8] and [21, Theorem 2.7]), and K. Kurdyka, who showed that this is also the case for *definable* $\mathscr{C}^1$ functions, that is, $\mathscr{C}^1$ functions whose graphs belong to o-minimal structures [22, Theorem 1]. (We refer to [23–25] for a comprehensive introduction to the theory of o-minimal structures.) More generally, $\mathscr{C}_{\mathrm{KL}}$ contains the $\mathscr{C}^1$ functions that are *tame* in the sense that their restrictions to open balls are definable in a same o-minimal structure [10, Proposition 6.2]. Tame functions behave well in two respects. First, their restrictions to line segments are piecewise-smooth and -monotone (see, e.g., [25, Section 2]). Second, the set of stationary

points of a tame $\mathscr{C}^1$ function either is discrete or contains a flat continuum [10, Proposition 6.5] and so cannot be both level-discrete and nondiscrete.

We conclude this section with examples of increasingly large o-minimal structures on the real field to illustrate the extent of $\mathscr{C}_{\mathrm{KL}}$.

The simplest o-minimal structure is $\mathbb{R}_{\mathrm{alg}} := (\mathcal{S}_n)_{n\in\mathbb{N}}$, where $\mathcal{S}_n$ is the class of semialgebraic sets in $\mathbb{R}^n$, that is, sets of the form

$$\bigcup_{i=1}^{k} \bigcap_{j=1}^{l} \big\{ \boldsymbol{x} \in \mathbb{R}^n : f_{ij}(\boldsymbol{x}) = 0,\ g_{ij}(\boldsymbol{x}) > 0 \big\}, \tag{8.8}$$

where the functions $f_{ij}, g_{ij} : \mathbb{R}^n \to \mathbb{R}$ are polynomials. Any o-minimal structure $(\mathcal{T}_n)$ is an expansion of $\mathbb{R}_{\mathrm{alg}}$ in the sense that $\mathcal{T}_n \supset \mathcal{S}_n$ for all n (in contrast, there is no "largest" expansion containing all o-minimal structures [26]).

The two other fundamental o-minimal structures are the structure $\mathbb{R}_{\mathrm{an}}$ of globally subanalytic sets [27] and the structure $\mathbb{R}_{\mathrm{an,exp}}$ of analytic-exponential sets [28]. The structure $\mathbb{R}_{\mathrm{an}}$ is the smallest o-minimal structure containing the graphs of the restricted analytic functions (that is, the functions $f : \mathbb{R}^n \to \mathbb{R}$ such that the restriction $f|_{[-1,1]^n}$ has an analytic extension on a neighborhood of $[-1,1]^n$ and $f|_{\mathbb{R}^n\setminus[-1,1]^n}$ is identically zero). The sets in $\mathbb{R}_{\mathrm{an}}$ can be described in a way similar to semialgebraic objects by letting the functions f_{ij} and g_{ij} in (8.8) be defined by composition from (i) polynomials, (ii) the restricted analytic functions, and (iii) the extended reciprocal function

$$\mathrm{inv} : t \in \mathbb{R} \longmapsto \begin{cases} 1/t & \text{if } t \neq 0, \\ 0 & \text{if } t = 0. \end{cases} \tag{8.9}$$

The structure $\mathbb{R}_{\mathrm{an}}$ is polynomially bounded [29] and hence does not define the infinite branches of the exponential function. The structure $\mathbb{R}_{\mathrm{an,exp}}$ is the smallest expansion of $\mathbb{R}_{\mathrm{an}}$ containing the graph of the exponential; it is obtained by also allowing composition from (iv) the exponential function and (v) the extended logarithm function $t \in \mathbb{R} \longmapsto \ln t$ if $t > 0$, 0 if $t \leqslant 0$.

Finally, note that some special functions such as the Riemann zeta function, the gamma function, and the error function are not definable in $\mathbb{R}_{\mathrm{an,exp}}$ [30]. The zeta and gamma functions are separately definable in two larger o-minimal structures constructed by taking the Pfaffian closures [31] of expansions of $\mathbb{R}_{\mathrm{alg}}$ with special power series [32, 33]. These structures also define the error function, but the existence of a further o-minimal expansion defining both the zeta and gamma functions remains a conjecture [34].

9. The case of tame subanalytic objectives

Here we consider the special case where the objective f is tame in $\mathbb{R}_{\mathrm{an}}$. We make the following definition.

Definition 9.1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *tame subanalytic* if its restrictions to open balls are definable in $\mathbb{R}_{\mathrm{an}}$.

The motivation is twofold. First, tame subanalyticity ensures that f has the classical Łojasiewicz property at every stationary point (see Definition 9.2 below). This will allow us to relate the convergence rate of $\mathscr{C}^0$-QMM algorithms to the geometry of the graph of f around attractors. Second, the class of tame subanalytic functions covers many practical objectives. The reason is that tameness is a

local definability property, making it possible to use infinite branches of analytic functions that are not definable in $\mathbb{R}_{\text{an}}$ (for example, the exponential function is tame subanalytic).

Definition 9.2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. We say that f has the *Łojasiewicz property* at $\boldsymbol{x}$ (or that f satisfies the *Łojasiewicz inequality* at $\boldsymbol{x}$) if there is an integer $\sigma \geqslant 2$ such that f satisfies the KL inequality at $\boldsymbol{x}$ with $\psi(t) \propto t^{1/\sigma}$.

The next result is a corollary of the Łojasiewicz inequality for subanalytic $\mathscr{C}^1$ functions [22]. We denote by $\mathscr{C}^1_{\text{TS}}$ the class of tame subanalytic $\mathscr{C}^1$ functions on $\mathbb{R}^n$.

Theorem 9.3. *A function $f \in \mathscr{C}^1_{\text{TS}}$ has the Łojasiewicz property at every point $\boldsymbol{x} \in \mathcal{S}_f$.*

Proof. Let $\boldsymbol{x} \in \mathcal{S}_f$ (assuming $\mathcal{S}_f \neq \emptyset$). The restriction $f|_{\mathscr{B}(\boldsymbol{x},1)}$ is $\mathbb{R}_{\text{an}}$-definable and thus subanalytic. Let $f_{\boldsymbol{x}} := f - f(\boldsymbol{x})$. Theorem ŁI in [22] tells us that there are positive constants c and ρ and an integer $\sigma \geqslant 2$ such that

$$\|\nabla f(\boldsymbol{y})\| \;\geqslant\; c\big(f_{\boldsymbol{x}}(\boldsymbol{y})\big)^{1-1/\sigma} \quad \text{for all } \; \boldsymbol{y} \in \mathscr{B}(\boldsymbol{x},1) \cap \{\, f_{\boldsymbol{x}} \in (0,\rho)\,\},$$

which shows that f has the KL property at $\boldsymbol{x}$ with $\psi(t) = (\sigma/c)t^{1/\sigma}$. □

The Łojasiewicz property means that $|f - f(\boldsymbol{x})|^{1-1/\sigma}\|\nabla f\|^{-1}$ is bounded around $\boldsymbol{x}$ for some integer $\sigma \geqslant 2$. The smallest such integer is called the *desingularizing exponent* of f at $\boldsymbol{x}$ and is also denoted by σ for simplicity. (The rational number $1 - 1/\sigma$ is known as the *Łojasiewicz exponent*.) The desingularizing exponent is a local flatness measure: the larger σ, the steeper the desingularizing function ψ near zero, and hence the flatter f around $\boldsymbol{x}$. This suggests that the convergence rate is inversely related to σ, as we now show.

Theorem 9.4. *Let $\{\boldsymbol{x}_p\}$ be a $\mathscr{C}^0$-QMM sequence with differentiable objective f. Assume that $\{\boldsymbol{x}_p\}$ converges, let $\boldsymbol{x} := \lim_p \boldsymbol{x}_p$, and let*

$$R_p \;:=\; \sum_{k=p}^{\infty} \|\,\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\,\|. \tag{9.1}$$

Assume also that f satisfies the Łojasiewicz inequality at $\boldsymbol{x}$.

(i) *If $\sigma = 2$ then there exists $\eta \in (0,1)$ such that $R_p = O(\eta^p)$.*

(ii) *If $\sigma \geqslant 3$ then $R_p = O\big(p^{-1/(\sigma-2)}\big)$.*

Proof. The limit $\boldsymbol{x}$ is a stationary point by Theorem 5.2(i). Suppose $\{\boldsymbol{x}_p\} \cap \mathcal{S}_f$ is empty (otherwise, by Corollary 4.3, R_p is eventually zero and we are done). By the Łojasiewicz property, we have

$$t_p^{1-1/\sigma} \;=\; O\big(\|\nabla f(\boldsymbol{x}_p)\|\big), \qquad t_p := f(\boldsymbol{x}_p) - f(\boldsymbol{x}).$$

From (8.6) and (8.7), we have, respectively,

$$\|\nabla f(\boldsymbol{x}_p)\| \;=\; O\big(\|\,\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\,\|\big) \quad \text{and} \quad R_p \;=\; O\big(t_p^{1/\sigma}\big).$$

Therefore, since $\|\,\boldsymbol{x}_{p+1} - \boldsymbol{x}_p\,\| = R_p - R_{p+1}$,

$$R_p^{\sigma-1} \;=\; O\big(t_p^{1-1/\sigma}\big) \;=\; O\big(\|\nabla f(\boldsymbol{x}_p)\|\big) \;=\; O(R_p - R_{p+1}).$$

Suppose that $\sigma = 2$. Then $R_{p+1} \leqslant R_p = O(R_p - R_{p+1})$, so there exists $\eta \in (0,1)$ such that $R_{p+1} \leqslant \eta R_p$ for sufficiently large p, and hence $R_p = O(\eta^p)$. Suppose now that $\sigma \geqslant 3$. Then there exist $c > 0$ and $q \in \mathbb{N}$ such that for all $p \geqslant q$,

$$c \leqslant (R_p - R_{p+1}) R_p^{1-\sigma} \leqslant \int_{R_{p+1}}^{R_p} t^{1-\sigma}\,\mathrm{d}t = \frac{1}{\sigma-2}\big(R_{p+1}^{2-\sigma} - R_p^{2-\sigma}\big).$$

Letting $c' := c(\sigma-2)$, we have for $p > q$ that

$$c'(p-q) \leqslant \sum_{k=q}^{p-1} \big(R_{k+1}^{2-\sigma} - R_k^{2-\sigma}\big) \leqslant R_p^{2-\sigma},$$

and so $R_p \leqslant \big(c'(p-q)\big)^{-1/(\sigma-2)} = O\big(p^{-1/(\sigma-2)}\big)$. □

Corollary 9.5. *Let $\{\boldsymbol{x}_p\}$ be a bounded $\mathscr{C}^0$-QMM sequence with objective $f \in \mathscr{C}^1_{\mathrm{TS}}$, let $\boldsymbol{x} := \lim_p \boldsymbol{x}_p$, and let σ be the desingularizing exponent of f at $\boldsymbol{x}$.*

(i) *If $\sigma = 2$ then there exists $\eta \in (0,1)$ such that $\|\boldsymbol{x}_p - \boldsymbol{x}\| = O(\eta^p)$.*

(ii) *If $\sigma \geqslant 3$ then $\|\boldsymbol{x}_p - \boldsymbol{x}\| = O\big(p^{-1/(\sigma-2)}\big)$.*

In other words, $\{\boldsymbol{x}_p\}$ converges R-linearly if $\sigma = 2$ and R-sublinearly otherwise.

Proof. By Theorems 8.4 and 9.3, $\{\boldsymbol{x}_p\}$ converges to a stationary point $\boldsymbol{x}$ and f has the Łojasiewicz property at this point. So Theorem 9.4 applies and the result follows from the fact that, for all p,

$$\|\boldsymbol{x}_p - \boldsymbol{x}\| = \lim_q \|\boldsymbol{x}_p - \boldsymbol{x}_q\| = \lim_q \left\| \sum_{k=p}^{q-1} (\boldsymbol{x}_k - \boldsymbol{x}_{k+1}) \right\| \leqslant R_p. \tag{9.2}$$ □

10. The case of $\mathscr{C}^2$ objectives

We now focus on the convergence rate of a $\mathscr{C}^0$-QMM sequence $\{\boldsymbol{x}_p\}$ whose objective is twice continuously differentiable around the limit $\boldsymbol{x}$. Theorem 10.2 below shows that $\{\boldsymbol{x}_p\}$ converges R-linearly if the sequence of the unit vectors in the directions of $\boldsymbol{x}_p - \boldsymbol{x}$ has no limit point in the null space of $\nabla^2 f(\boldsymbol{x})$ (the Hessian at $\boldsymbol{x}$), or, in other words, if the trajectory of the iterates does not hug the null space of $\nabla^2 f(\boldsymbol{x})$ too tightly. A consequence is that $\{\boldsymbol{x}_p\}$ converges R-linearly if $\boldsymbol{x}$ is an isolated stationary point and $\nabla^2 f(\boldsymbol{x})$ is nonzero (Corollary 10.3).

We need the following Lemma.

Lemma 10.1. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix. There are positive constants c and c' such that for all $\boldsymbol{y} \in \mathbb{R}^n$,*

$$c\,\|\boldsymbol{A}\boldsymbol{y}\| \leqslant \operatorname{dist}(\boldsymbol{y}, \operatorname{null}(\boldsymbol{A})) \leqslant c'\,\|\boldsymbol{A}\boldsymbol{y}\|. \tag{10.1}$$

Proof. Given a subspace $\mathcal{S}$ of $\mathbb{R}^n$, we let $\boldsymbol{P}_{\mathcal{S}} \in \mathbb{R}^{n\times n}$ denote the orthogonal projection onto $\mathcal{S}$. Since $\boldsymbol{A}$ is symmetric, the orthogonal complement of $\operatorname{null}(\boldsymbol{A})$ is $\operatorname{ran}(\boldsymbol{A})$ (the range of $\boldsymbol{A}$), and thus

$$\operatorname{dist}(\boldsymbol{y}, \operatorname{null}(\boldsymbol{A})) = \|\boldsymbol{y} - \boldsymbol{P}_{\operatorname{null}(\boldsymbol{A})}(\boldsymbol{y})\| = \|\boldsymbol{P}_{\operatorname{ran}(\boldsymbol{A})}(\boldsymbol{y})\|. \tag{10.2}$$

The matrix $\boldsymbol{A}$ can be factored as $\boldsymbol{A} = \boldsymbol{V}\operatorname{diag}(\lambda_1, \ldots, \lambda_n)\boldsymbol{V}^T$, where $\boldsymbol{V}$ is an orthogonal matrix whose columns $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are eigenvectors of $\boldsymbol{A}$, and where $\lambda_1, \ldots, \lambda_n$ are the corresponding eigenvalues. Let $m := \operatorname{rank}(\boldsymbol{A})$. If $m = 0$, the result is

trivial, so assume that $m \geqslant 1$. Without loss of generality, assume also that $\lambda_i \neq 0$ for all $i = 1, \ldots, m$. Then for all $\boldsymbol{y} \in \mathbb{R}^n$ we have

$$\|\boldsymbol{A}\boldsymbol{y}\|^2 = \sum_{i=1}^{m} \lambda_i^2 (\boldsymbol{v}_i^T \boldsymbol{y})^2 \quad \text{and} \quad \|\boldsymbol{P}_{\mathrm{ran}(\boldsymbol{A})}(\boldsymbol{y})\|^2 = \sum_{i=1}^{m} (\boldsymbol{v}_i^T \boldsymbol{y})^2. \tag{10.3}$$

From (10.2) and (10.3) it follows that

$$\Big(\max_{i=1,\ldots,m} \lambda_i^2\Big)^{-1} \|\boldsymbol{A}\boldsymbol{y}\|^2 \leqslant \operatorname{dist}(\boldsymbol{y}, \operatorname{null}(\boldsymbol{A}))^2 \leqslant \Big(\min_{i=1,\ldots,m} \lambda_i^2\Big)^{-1} \|\boldsymbol{A}\boldsymbol{y}\|^2. \qquad \square$$

Theorem 10.2. *Let $\{\boldsymbol{x}_p\}$ be a $\mathscr{C}^0$-QMM sequence with differentiable objective f. Suppose that $\{\boldsymbol{x}_p\}$ converges and let $\boldsymbol{x} := \lim_p \boldsymbol{x}_p$. Suppose also that f is $\mathscr{C}^2$ in a neighborhood of $\boldsymbol{x}$ and that $\boldsymbol{x}_p \neq \boldsymbol{x}$ for all p. If*

$$\mathcal{L}_{\{\boldsymbol{u}_p\}} \cap \operatorname{null}(\nabla^2 f(\boldsymbol{x})) = \emptyset, \qquad \boldsymbol{u}_p := \frac{\boldsymbol{x}_p - \boldsymbol{x}}{\|\boldsymbol{x}_p - \boldsymbol{x}\|}, \tag{10.4}$$

then $\{\boldsymbol{x}_p\}$ converges R-linearly.

Proof. The limit $\boldsymbol{x}$ is a stationary point by Theorem 5.2(i). Using the second-order Taylor-Lagrange formula for f, we find that there is an $r > 0$ such that for all $\boldsymbol{y} \in \mathscr{B}(\boldsymbol{x}, r)$ we have

$$|f(\boldsymbol{y}) - f(\boldsymbol{x})| \leqslant \tfrac{1}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \sup_{t \in [0,1]} \|\nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))\|.$$

Hence, since $\{f(\boldsymbol{x}_p)\}$ is decreasing,

$$f(\boldsymbol{x}_p) - f(\boldsymbol{x}) = O\big(\|\boldsymbol{x}_p - \boldsymbol{x}\|^2\big).$$

Furthermore, from the proof of Theorem 9.4, $\{\boldsymbol{x}_p\}$ converges R-linearly if

$$\big(f(\boldsymbol{x}_p) - f(\boldsymbol{x})\big)^{1/2} = O\big(\|\nabla f(\boldsymbol{x}_p)\|\big).$$

Thus, it suffices to show that

$$\|\boldsymbol{x}_p - \boldsymbol{x}\| = O\big(\|\nabla f(\boldsymbol{x}_p)\|\big),$$

or, equivalently,

$$\liminf_{p \to \infty} \frac{\|\nabla f(\boldsymbol{x}_p)\|}{\|\boldsymbol{x}_p - \boldsymbol{x}\|} > 0.$$

By the differentiability of ∇f at $\boldsymbol{x}$, we have for all $\boldsymbol{y} \in \mathbb{R}^n$ that

$$\nabla f(\boldsymbol{y}) = \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \|\boldsymbol{y} - \boldsymbol{x}\| \boldsymbol{\varepsilon}(\boldsymbol{y} - \boldsymbol{x}),$$

where $\boldsymbol{\varepsilon}(\boldsymbol{h})$ goes to zero as $\boldsymbol{h} \to \boldsymbol{0}$. Therefore

$$\|\nabla f(\boldsymbol{x}_p)\| \geqslant \|\boldsymbol{x}_p - \boldsymbol{x}\| \big(\|\nabla^2 f(\boldsymbol{x}) \boldsymbol{u}_p\| - \|\boldsymbol{\varepsilon}(\boldsymbol{x}_p - \boldsymbol{x})\|\big)$$

and hence

$$\liminf_{p \to \infty} \frac{\|\nabla f(\boldsymbol{x}_p)\|}{\|\boldsymbol{x}_p - \boldsymbol{x}\|} \geqslant \liminf_{p \to \infty} \|\nabla^2 f(\boldsymbol{x}) \boldsymbol{u}_p\|.$$

By Lemma 10.1, $\liminf_p \|\nabla^2 f(\boldsymbol{x}) \boldsymbol{u}_p\| > 0$ if and only if

$$\liminf_{p \to \infty} \operatorname{dist}\big(\boldsymbol{u}_p, \operatorname{null}(\nabla^2 f(\boldsymbol{x}))\big) > 0,$$

which in turn is equivalent to (10.4), completing the proof. $\square$

Corollary 10.3. *Let $\{\boldsymbol{x}_p\}$ be a $\mathscr{C}^0$-QMM sequence with differentiable objective f. Assume that $\{\boldsymbol{x}_p\}$ converges and that f is $\mathscr{C}^2$ in a neighborhood of $\boldsymbol{x} := \lim_p \boldsymbol{x}_p$. If $\boldsymbol{x}$ is isolated in $\mathcal{S}_f$ and $\nabla^2 f(\boldsymbol{x}) \neq \boldsymbol{0}$, then $\{\boldsymbol{x}_p\}$ converges R-linearly.*

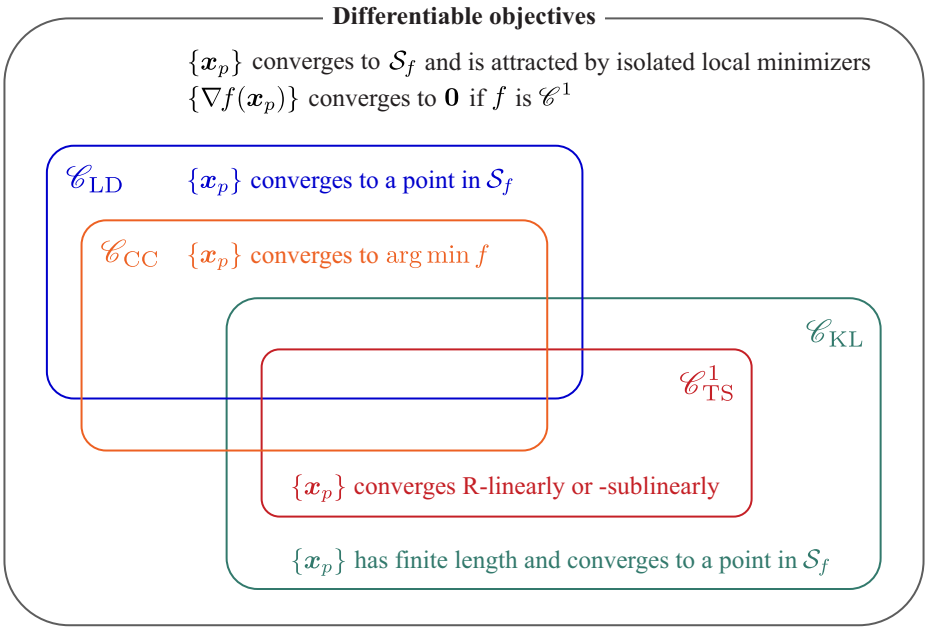

FIGURE 2. Convergence of bounded $\mathscr{C}^0$-QMM sequences with differentiable objectives (Theorems 5.2 and 6.2, Corollary 6.3, Theorems 7.3 and 8.4, and Corollary 9.5).

Proof. Suppose that $\nabla^2 f(\boldsymbol{x}) \neq \mathbf{0}$. Then $\nabla^2 f(\cdot)$ is nonzero around $\boldsymbol{x}$ and so the null space of $\nabla^2 f(\boldsymbol{x})$ is equal to the tangent space of $\mathcal{S}_f$ at $\boldsymbol{x}$. It follows from Theorem 10.2 that $\{\boldsymbol{x}_p\}$ converges R-linearly if no limit point of $\{\boldsymbol{u}_p\}$ is tangent to $\mathcal{S}_f$ at $\boldsymbol{x}$, meaning that $\mathcal{S}_f$ does not contain any $\mathscr{C}^1$ curve $\omega : (-a, a) \to \mathbb{R}^n$, $a > 0$, such that $\omega(0) = \boldsymbol{x}$ and $\omega'(0) \in \mathcal{L}_{\{\boldsymbol{u}_p\}}$. This is trivially the case when $\boldsymbol{x}$ is isolated in $\mathcal{S}_f$. □

11. SUMMARY OF THE CONVERGENCE RESULTS

The convergence properties proved in Sections 5–9 are summarized in Figure 2, where $\mathscr{C}_{\mathrm{LD}}$ denotes the subclass of functions whose set of stationary points is level-discrete (Definition 6.1), and where $\mathscr{C}_{\mathrm{CC}}$ is the subclass of convex and coercive functions. Recall that $\mathscr{C}_{\mathrm{KL}}$ is the subclass of functions satisfying the Kurdyka-Łojasiewicz inequality at all points (Definition 8.1) and $\mathscr{C}^1_{\mathrm{TS}}$ is the subclass of tame subanalytic $\mathscr{C}^1$ functions (Definition 9.1). Example showing that the sets $\mathscr{C}^1_{\mathrm{TS}} \setminus \mathscr{C}_{\mathrm{LD}}$, $\mathscr{C}^1 \cap \mathscr{C}_{\mathrm{KL}} \setminus \mathscr{C}_{\mathrm{LD}} \setminus \mathscr{C}^1_{\mathrm{TS}}$, and $\mathscr{C}^1 \cap \mathscr{C}_{\mathrm{LD}} \cap \mathscr{C}_{\mathrm{KL}} \setminus \mathscr{C}^1_{\mathrm{TS}}$ are nonempty are given in [10].

It is important to emphasize that these results also hold when the objective f is restricted to an open set $\mathcal{O}$ provided $\{\boldsymbol{x}_p\}$ remains in a compact set contained in $\mathcal{O}$. Since we assume that $\{\boldsymbol{x}_p\}$ is bounded, this is equivalent to imposing that $\inf_p \operatorname{dist}(\boldsymbol{x}_p, \mathbb{R}^n \setminus \mathcal{O}) > 0$. A sufficient (but not necessary) condition for this to happen is that the starting point $\boldsymbol{x}_0$ is in a basin of f contained in $\mathcal{O}$.

The convergence to a stationary point $\boldsymbol{x}$ is R-linear if $\boldsymbol{x}$ is isolated in $\mathcal{S}_f$ and f is $\mathscr{C}^2$ around $\boldsymbol{x}$ with $\nabla^2 f(\boldsymbol{x}) \neq \mathbf{0}$ (Corollary 10.3). In particular, the convergence rate is R-linear if f is $\mathscr{C}^2$ around $\boldsymbol{x}$ and $\nabla^2 f(\boldsymbol{x})$ is nonsingular.

We now highlight how our results complement those relating to MM in the general context of nonsmooth optimization.

First, it comes as no surprise that the convergence properties are weak when f is only assumed to be directionally differentiable: it is shown in [12] (respectively, [16]) that any limit point $\boldsymbol{x}$ of a block-coordinate exact MM sequence (respectively, an inexact MM sequence) is stationary in the usual sense that the directional derivatives $\nabla_{\boldsymbol{v}} f(\boldsymbol{x}) := \lim_{t\to 0^+}(1/t)(f(\boldsymbol{x}+t\boldsymbol{v}) - f(\boldsymbol{x}))$ are all nonnegative. A slightly stronger convergence guarantee is given in [14] for exact MM sequences whose approximation errors $f(\cdot\,|\,\boldsymbol{x}_p) - f$ are L-smooth with L independent of p (a function g is said to be L-smooth if it is differentiable and ∇g is L-Lipschitz continuous); such sequences are asymptotically stationary in the sense that $\liminf_p \inf_{\|v\|=1} \nabla_{\boldsymbol{v}} f(\boldsymbol{x}_p) \geqslant 0$. However, this does not exclude the possibility that $\{\boldsymbol{x}_p\}$ diverges.

Second, the convergence results for nonsmooth MM do not cover all differentiable objectives or even one of the subclasses $\mathscr{C}^1_{\mathrm{LD}}$, $\mathscr{C}^1_{\mathrm{CC}}$ or $\mathscr{C}^1_{\mathrm{TS}}$. Theorem 4.1 in [13] ensures finite-length convergence of $\mathscr{C}^0$-QMM provided that f is coercive, L-smooth, and subanalytic. Similarly, Theorem 8.4 is a special case of Theorem 3.2 in [11] if f is L-smooth on a compact set containing $\{\boldsymbol{x}_p\}$ for sufficiently small L. Under this assumption, Theorem 6 in [15] yields the same convergence rates as in Corollary 9.5, though only for exact QMM.

Finally, note that in [12, 14, 15] the inner optimizations problems are assumed to be solved exactly without questioning their feasibility (be it theoretical or computational), while the studies in [11, 13, 16] do not address the problem of generating iterates satisfying the assumptions made. In constrast, proper $\mathscr{C}^0$-QMM sequences can be generated efficiently with good stability properties, as we show in the next sections.

12. Large-scale implementation

12.1. **Generating QMM sequences using truncated CG.** We begin by motivating the use of the CG method to compute each iterate of a QMM sequence.

Recall from (3.5) that the exact QMM map $\boldsymbol{\Phi}_0 : \boldsymbol{x} \longmapsto \{\phi(\boldsymbol{x})\}$ is defined as follows: $\phi(\boldsymbol{x})$ is the global minimizer of the surrogate function $f(\cdot\,|\,\boldsymbol{x})$, which is the solution to the system

$$\boldsymbol{W}(\boldsymbol{x})\boldsymbol{y} \;=\; \boldsymbol{v}(\boldsymbol{x}), \qquad \boldsymbol{v}(\boldsymbol{x}) := \boldsymbol{W}(\boldsymbol{x})\boldsymbol{x} - \boldsymbol{w}(\boldsymbol{x}). \tag{12.1}$$

The next lemma defines $\boldsymbol{\Phi}_\gamma$ in terms of an upper bound on the energy norm of the error $E(\cdot\,|\,\boldsymbol{x})$ defined by

$$E(\boldsymbol{y}\,|\,\boldsymbol{x}) \;:=\; \|\boldsymbol{y} - \phi(\boldsymbol{x})\|_{\boldsymbol{W}(\boldsymbol{x})}. \tag{12.2}$$

Lemma 12.1. *Let $\gamma \in (0,1)$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then $\boldsymbol{y} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x})$ if and only if*

$$E(\boldsymbol{y}\,|\,\boldsymbol{x}) \;\leqslant\; \sqrt{\gamma}\, E(\boldsymbol{x}\,|\,\boldsymbol{x}). \tag{12.3}$$

Proof. For any $\boldsymbol{z} \in \mathbb{R}^n$ we have

$$\begin{aligned} E(\boldsymbol{z}\,|\,\boldsymbol{x})^2 \;&=\; \|\boldsymbol{z} - \boldsymbol{x} + \boldsymbol{W}(\boldsymbol{x})^{-1}\boldsymbol{w}(\boldsymbol{x})\|^2_{\boldsymbol{W}(\boldsymbol{x})} \\ &=\; \|\boldsymbol{z} - \boldsymbol{x}\|^2_{\boldsymbol{W}(\boldsymbol{x})} + 2(\boldsymbol{z} - \boldsymbol{x})^T\boldsymbol{w}(\boldsymbol{x}) + \|\boldsymbol{w}(\boldsymbol{x})\|^2_{\boldsymbol{W}(\boldsymbol{x})^{-1}} \\ &=\; 2\big(f(\boldsymbol{z}\,|\,\boldsymbol{x}) - f(\boldsymbol{x})\big) + \|\boldsymbol{w}(\boldsymbol{x})\|^2_{\boldsymbol{W}(\boldsymbol{x})^{-1}} \\ &=\; 2\big(f(\boldsymbol{z}\,|\,\boldsymbol{x}) - \min f(\cdot\,|\,\boldsymbol{x})\big), \end{aligned}$$

where the last equality follows from (4.2). Hence (12.3) is equivalent to

$$f(\boldsymbol{y}\,|\,\boldsymbol{x}) - \min f(\cdot\,|\,\boldsymbol{x}) \;\leqslant\; \gamma\big(f(\boldsymbol{x}) - \min f(\cdot\,|\,\boldsymbol{x})\big) \iff \boldsymbol{y} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x}). \qquad \square$$

It follows that each iteration of a QMM algorithm consists of solving the following inner optimization problem: given $\boldsymbol{x} := \boldsymbol{x}_p$, find an approximate solution to the system (12.1) within the error bound (12.3), for some contraction number γ independent of p. In practical applications of QMM, the weighting matrices $\boldsymbol{W}(\boldsymbol{x})$ are usually large and sparse, and since these matrices are symmetric positive definite, CG is the natural choice for the inner solver.

The next result shows that QMM sequences can be generated by using truncated CG. We assume the use of a preconditioner: given a nonsingular matrix $\boldsymbol{H}(\boldsymbol{x}) \in \mathbb{R}^{n\times n}$, the preconditioned inner system has the form

$$\underbrace{\boldsymbol{H}(\boldsymbol{x})^{-1}\boldsymbol{W}(\boldsymbol{x})\boldsymbol{H}(\boldsymbol{x})^{-T}}_{=:\;\boldsymbol{W}'(\boldsymbol{x})}\boldsymbol{y}' \;=\; \boldsymbol{H}(\boldsymbol{x})^{-1}\boldsymbol{v}(\boldsymbol{x}), \qquad \boldsymbol{y}' := \boldsymbol{H}(\boldsymbol{x})^T\boldsymbol{y}. \tag{12.4}$$

We define $\boldsymbol{\Psi}(\boldsymbol{x}) := \{\boldsymbol{y}_j(\boldsymbol{x})\}_{j\geqslant 1}$, where $\boldsymbol{y}_j(\boldsymbol{x})$ is the approximate solution to (12.1) obtained after j steps of the preconditioned conjugate gradient (PCG) algorithm starting from $\boldsymbol{y}_0 := \boldsymbol{x}$ (equivalently, $\boldsymbol{H}(\boldsymbol{x})^T\boldsymbol{y}_j(\boldsymbol{x})$ is the jth iterate of unpreconditioned CG for (12.4) starting from $\boldsymbol{H}(\boldsymbol{x})^T\boldsymbol{x}$). For simplicity, we will omit the dependence of $\boldsymbol{y}_j$ on $\boldsymbol{x}$ when there is no ambiguity.

Theorem 12.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and suppose $\boldsymbol{W}$ is continuous. Any bounded sequence $\{\boldsymbol{x}_p\}$ such that $\boldsymbol{x}_{p+1} \in \boldsymbol{\Psi}(\boldsymbol{x}_p)$ for all p is a QMM sequence.*

Proof. Let $\phi'(\boldsymbol{x})$ be the solution to the preconditioned system and let $E'(\cdot\,|\,\boldsymbol{x})$ be the corresponding error energy norm. For any $\boldsymbol{y}' \in \mathbb{R}^n$ we have

$$\begin{aligned} E'(\boldsymbol{y}'\,|\,\boldsymbol{x}) &:= \|\boldsymbol{y}' - \phi'(\boldsymbol{x})\|_{\boldsymbol{W}'(\boldsymbol{x})} \\ &= \|\boldsymbol{H}(\boldsymbol{x})^T(\boldsymbol{y} - \phi(\boldsymbol{x}))\|_{\boldsymbol{H}(\boldsymbol{x})^{-1}\boldsymbol{W}(\boldsymbol{x})\boldsymbol{H}(\boldsymbol{x})^{-T}} \\ &= \big((\boldsymbol{y} - \phi(\boldsymbol{x}))^T\boldsymbol{W}(\boldsymbol{x})(\boldsymbol{y} - \phi(\boldsymbol{x}))\big)^{1/2} \\ &= E(\boldsymbol{y}\,|\,\boldsymbol{x}). \end{aligned} \tag{12.5}$$

Let $j \geqslant 1$ and let $\kappa'(\boldsymbol{x})$ denote the spectral condition number of $\boldsymbol{W}'(\boldsymbol{x})$. From Theorem 3.1.1 in [35] it follows that

$$E(\boldsymbol{y}_j\,|\,\boldsymbol{x}) \;\leqslant\; g_j(h(\boldsymbol{x}))\,E(\boldsymbol{x}\,|\,\boldsymbol{x}), \tag{12.6a}$$

where $h : \mathbb{R}^n \to [0,1]$ and $g_j : [0,1] \to [0,1]$ are defined by

$$h(\boldsymbol{x}) \;:=\; \frac{\kappa'(\boldsymbol{x})^{1/2} - 1}{\kappa'(\boldsymbol{x})^{1/2} + 1} \quad\text{and}\quad g_j(t) \;:=\; \frac{2t^j}{1 + t^{2j}}. \tag{12.6b}$$

Suppose $\{\boldsymbol{x}_p\}$ is bounded and let $\mathcal{C}$ be a compact set containing $\{\boldsymbol{x}_p\}$. From the continuity of $\boldsymbol{W}$ we have $\sup_{\mathcal{C}} \kappa' < +\infty$, and hence $\sup_{\mathcal{C}} h < 1$. Since $g_j(t)$ increases with increasing t and decreases with increasing j, we have for all p that

$$g_j(h(\boldsymbol{x}_p)) \;\leqslant\; g_j(\sup_{\mathcal{C}} h) \;\leqslant\; g_1(\sup_{\mathcal{C}} h) \;<\; 1.$$

Let $\gamma := (g_1(\sup_{\mathcal{C}} h))^2$. Using Lemma 12.1, we deduce that $\boldsymbol{y}_j(\boldsymbol{x}_p) \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x}_p)$ and the result follows. $\square$

Algorithm 1 PCG for the inner system $\boldsymbol{W}(\boldsymbol{x})\boldsymbol{y} = \boldsymbol{v}(\boldsymbol{x})$.

INPUT: starting point $\boldsymbol{y}_0$.
$\boldsymbol{r}_0 = \boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{W}(\boldsymbol{x})\boldsymbol{y}_0$
$\boldsymbol{p}_0 = \boldsymbol{s}_0 = \boldsymbol{M}(\boldsymbol{x})^{-1}\boldsymbol{r}_0$
$\beta_0 = \boldsymbol{r}_0^T\boldsymbol{s}_0$
for $j = 0, 1, 2, \ldots$ **do**
  $\boldsymbol{q}_j = \boldsymbol{W}(\boldsymbol{x})\boldsymbol{p}_j$
  $\alpha_j = \beta_j / (\boldsymbol{p}_j^T\boldsymbol{q}_j)$
  $\boldsymbol{y}_{j+1} = \boldsymbol{y}_j + \alpha_j\boldsymbol{p}_j$
  $\boldsymbol{r}_{j+1} = \boldsymbol{r}_j - \alpha_j\boldsymbol{q}_j$
  $\boldsymbol{s}_{j+1} = \boldsymbol{M}(\boldsymbol{x})^{-1}\boldsymbol{r}_{j+1}$
  $\beta_{j+1} = \boldsymbol{r}_{j+1}^T\boldsymbol{s}_{j+1}$
  $\boldsymbol{p}_{j+1} = \boldsymbol{s}_{j+1} + (\beta_{j+1}/\beta_j)\boldsymbol{p}_j$
end for

The PCG method is given in Algorithm 1. Ideally, the preconditioner

$$\boldsymbol{M}(\boldsymbol{x}) := \boldsymbol{H}(\boldsymbol{x})\boldsymbol{H}(\boldsymbol{x})^T \tag{12.7}$$

should be chosen so that (i) linear systems with coefficient matrix $\boldsymbol{H}(\boldsymbol{x})$ are easy to solve and (ii) $\boldsymbol{W}'(\boldsymbol{x})$ has a condition number close to one or properly distributed eigenvalues [35].

12.2. **Control of the contraction number.** We now describe an inexpensive stopping criterion for PCG that allows to control the contraction number of a QMM sequence.

Let j_1 and j_2 be nonnegative integers with $j_1 < j_2$. The j_1th and j_2th error energy norms are related by the identity

$$E(\boldsymbol{y}_{j_1}|\boldsymbol{x})^2 - E(\boldsymbol{y}_{j_2}|\boldsymbol{x})^2 = \sum_{i=j_1}^{j_2-1} \alpha_i\beta_i =: \zeta_{j_1,j_2}, \tag{12.8}$$

where α_i is the step length in the ith conjugate direction and β_i is the square of the weighted norm of the ith residual [36,37] (to simplify the notation, we omit the dependence of α_i and β_i on $\boldsymbol{x}$). Note that $\alpha_j\beta_j \geqslant 0$ with equality if and only if $\boldsymbol{r}_j = \boldsymbol{0}$, so the sequence $\{E(\boldsymbol{y}_j|\boldsymbol{x})\}$ is strictly decreasing until the residual is zero. Using (12.8) and Lemma 12.1, we find that $\boldsymbol{y}_j \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x})$ if and only if

$$(1-\gamma)\,E(\boldsymbol{y}_j|\boldsymbol{x})^2 \leqslant \gamma\zeta_{0,j}. \tag{12.9}$$

Let k be a positive integer. If the error energy norm decreases sufficiently between iterations j and $j+k$ (so that $E(\boldsymbol{y}_{j+k}|\boldsymbol{x})^2 \ll E(\boldsymbol{y}_j|\boldsymbol{x})^2$), it follows from (12.8) that

$$E(\boldsymbol{y}_j|\boldsymbol{x})^2 \approx \zeta_{j,j+k}. \tag{12.10}$$

Using this approximation in (12.9) and the fact that $\zeta_{0,j} + \zeta_{j,j+k} = \zeta_{0,j+k}$, we obtain the stopping criterion

$$\zeta_{j,j+k} \leqslant \gamma\zeta_{0,j+k}. \tag{12.11}$$

This test ensures that for sufficiently large k the QMM sequences covered by Theorem 12.2 have an actual contraction number close to γ (that is, $\boldsymbol{x}_{p+1} \in \boldsymbol{\Phi}_{\gamma'}(\boldsymbol{x}_p)$

for all p and some $\gamma' \approx \gamma$). We call the integer k the *stopping delay*. Because it takes $j+k$ iterations to test whether $\boldsymbol{y}_j \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x})$, we expect the actual contraction number to be smaller than γ in practice (see Section 12.4). Furthermore, the definition of $\zeta_{\cdot,\,j+k}$ involves only quantities that are available in the algorithm, so the cost of checking (12.11) is negligible.

12.3. **Numerical stability.** An important question is the numerical stability of CG-based QMM. In exact arithmetic, the direction vectors are mutually conjugate and the residuals are mutually orthogonal (that is, $\boldsymbol{p}_j^T \boldsymbol{W}(\boldsymbol{x}) \boldsymbol{p}_k = \boldsymbol{r}_j^T \boldsymbol{s}_k = 0$ if $j \neq k$). Consequently, $\boldsymbol{y}_j$ minimizes the error energy norm over the affine space

$$\boldsymbol{y}_0 \,+\, \operatorname{span}\big\{\, \boldsymbol{s}_0, \boldsymbol{M}(\boldsymbol{x})^{-1}\boldsymbol{W}(\boldsymbol{x})\,\boldsymbol{s}_0, \ldots, (\boldsymbol{M}(\boldsymbol{x})^{-1}\boldsymbol{W}(\boldsymbol{x}))^{j-1}\boldsymbol{s}_0 \,\big\}. \tag{12.12}$$

Recalling the relation (12.5), this property yields the following sharp upper bound [35, Section 3.1]:

$$\frac{E(\boldsymbol{y}_j\,|\,\boldsymbol{x})}{E(\boldsymbol{y}_0\,|\,\boldsymbol{x})} \;\leqslant\; \min_{P\,\in\,\mathscr{P}_{j,1}} \max_{i\,=\,1,\ldots,n} |P(\lambda_i'(\boldsymbol{x}))|, \tag{12.13}$$

where $\mathscr{P}_{j,1}$ is the set of polynomials of degree at most j with constant term 1, and $\lambda_1'(\boldsymbol{x}) \leqslant \cdots \leqslant \lambda_n'(\boldsymbol{x})$ are the eigenvalues of the preconditioned matrix $\boldsymbol{W}'(\boldsymbol{x})$. The error bound (12.6) is obtained by taking P to be the jth-degree Chebyshev polynomial shifted to the interval $[\lambda_1'(\boldsymbol{x}), \lambda_n'(\boldsymbol{x})]$ and normalized so that $P(0) = 1$. Thus Theorem 12.2 relies on the conjugacy of the direction vectors and the orthogonality of the residuals. However, these properties can be rapidly and completely lost in finite precision arithmetic (see, e.g., [38, Section 5] and [39, Section 5.1]), which calls into question the practical relevance of the theorem. Fortunately, we can remove this doubt.

The following results are immediate consequences of the stability analysis given by Greenbaum in [40, Theorems 1′ and 3′].

(i) The coefficients α_j and β_j generated by a finite precision PCG recurrence applied to the inner system $\boldsymbol{W}(\boldsymbol{x})\boldsymbol{y} = \boldsymbol{v}(\boldsymbol{x})$ are equal to those generated by an exact CG recurrence applied to a larger system $\overline{\boldsymbol{W}}(\boldsymbol{x})\overline{\boldsymbol{y}} = \overline{\boldsymbol{v}}(\boldsymbol{x})$. Furthermore, the eigenvalues of $\overline{\boldsymbol{W}}(\boldsymbol{x})$ all lie in the set

$$\overline{\Lambda}(\boldsymbol{x}) \;:=\; \bigcup_{i=1}^{n} \big[\lambda_i'(\boldsymbol{x}) - \tau,\, \lambda_i'(\boldsymbol{x}) + \tau\big], \tag{12.14}$$

where $\tau \ll \lambda_n'(\boldsymbol{x})$ depends on the machine precision. (Greenbaum gives an upper bound on τ that largely overestimates the size the intervals containing the eigenvalues of $\overline{\boldsymbol{W}}(\boldsymbol{x})$; empirical evidence suggests that τ can be chosen much smaller than the proven bound [41].)

(ii) Let $\overline{E}(\overline{\boldsymbol{y}}_j\,|\,\boldsymbol{x})$ denote the energy norm of the error of the jth iterate of the exact CG recurrence applied to the larger system. If τ is sufficiently small, then

$$\frac{E(\boldsymbol{y}_j\,|\,\boldsymbol{x})}{E(\boldsymbol{y}_0\,|\,\boldsymbol{x})} \;\leqslant\; \big(1 + O\big(\tau/\lambda_n'(\boldsymbol{x})\big)\big)\, \frac{\overline{E}(\overline{\boldsymbol{y}}_j\,|\,\boldsymbol{x})}{\overline{E}(\overline{\boldsymbol{y}}_0\,|\,\boldsymbol{x})}. \tag{12.15}$$

The ratio $\overline{E}(\overline{\boldsymbol{y}}_j\,|\,\boldsymbol{x})/\,\overline{E}(\overline{\boldsymbol{y}}_0\,|\,\boldsymbol{x})$ has an upper bound similar to (12.13), with $\lambda_1'(\boldsymbol{x})$, $\ldots, \lambda_n'(\boldsymbol{x})$ replaced by the eigenvalues of $\overline{\boldsymbol{W}}(\boldsymbol{x})$; so it follows that

$$\frac{E(\boldsymbol{y}_j\,|\,\boldsymbol{x})}{E(\boldsymbol{y}_0\,|\,\boldsymbol{x})} \;\leqslant\; \bigl(1 + O\bigl(\tau/\lambda_n'(\boldsymbol{x})\bigr)\bigr) \min_{P\,\in\,\mathscr{P}_{j,1}} \max_{\lambda\,\in\,\overline{\Lambda}(\boldsymbol{x})} |P(\lambda)|. \tag{12.16}$$

Hence if the smallest eigenvalue of $\boldsymbol{W}'(\boldsymbol{x})$ satisfies $(\lambda_1'(\boldsymbol{x}) - \tau)/\lambda_1'(\boldsymbol{x}) \approx 1$, then the error bound (12.13)—and hence the Chebyshev bound (12.6)—holds to a close approximation in finite precision arithmetic. We conclude that Theorem 12.2 holds in finite precision arithmetic under the additional assumption that $\sup_{\mathcal{B}} \lambda_1'$ is not too small, which can be ensured by proper preconditioning.

Turning to the stability of the inner stopping criterion (12.11), we need to check whether $\zeta_{j,\,j+k} := \sum_{i=j}^{j+k-1} \alpha_i\,\beta_i$ is a reliable estimate of $E(\boldsymbol{y}_j\,|\,\boldsymbol{x})^2$ in finite precision arithmetic. Let $m(\boldsymbol{x})$ be the maximum number of nonzeros per row of $\boldsymbol{W}(\boldsymbol{x})$, and let $\kappa(\boldsymbol{x})$ and $\kappa^\dagger(\boldsymbol{x})$ denote the spectral condition numbers of $\boldsymbol{W}(\boldsymbol{x})$ and $\boldsymbol{M}(\boldsymbol{x})$, respectively. Suppose that

$$\bigl(n + m(\boldsymbol{x})\,n^{1/2}\bigr)\kappa(\boldsymbol{x})\mathbf{u} \;\ll\; 1 \quad\text{and}\quad n^2\kappa^\dagger(\boldsymbol{x})\mathbf{u} \;\ll\; 1, \tag{12.17}$$

where $\mathbf{u}$ is the unit roundoff. From Theorem 4.4 in [37] the iterates and the scalar quantities generated by the PCG algorithm in finite precision arithmetic satisfy

$$E(\boldsymbol{y}_j\,|\,\boldsymbol{x})^2 - E(\boldsymbol{y}_{j+k}\,|\,\boldsymbol{x})^2 \;=\; \zeta_{j,\,j+k}(1 + \dot{\varepsilon}_{j,k}) \;+\; E(\boldsymbol{y}_j\,|\,\boldsymbol{x})\ddot{\varepsilon}_{j,k} + O(\mathbf{u}^2), \tag{12.18}$$

where $\dot{\varepsilon}_{j,k}$ and $\ddot{\varepsilon}_{j,k}$ account for rounding errors and depend on $\boldsymbol{x}$. The first error term, $\dot{\varepsilon}_{j,k}$, is negligible, and

$$|\ddot{\varepsilon}_{j,k}| \;\leqslant\; \kappa(\boldsymbol{x})^{1/2}\,P(n,k)\,Q_j(\boldsymbol{x})\mathbf{u} + O(\mathbf{u}^2), \tag{12.19}$$

where P is a small degree polynomial in n and k with constant coefficients and

$$Q_j(\boldsymbol{x}) \;:=\; \|\boldsymbol{W}(\boldsymbol{x})\|^{1/2}\bigl(\|\phi(\boldsymbol{x})\| + \max_{i\,=\,0,\ldots,j+1} \|\boldsymbol{y}_i\|\bigr). \tag{12.20}$$

It follows that if the error energy norm decreases sufficiently between iterations j and $j + k$, then

$$\bigl|E(\boldsymbol{y}_j\,|\,\boldsymbol{x})^2 - \zeta_{j,\,j+k}\bigr| \;\lesssim\; E(\boldsymbol{y}_j\,|\,\boldsymbol{x})\,\kappa(\boldsymbol{x})^{1/2}\,P(n,k)\,Q_j(\boldsymbol{x})\mathbf{u}, \tag{12.21}$$

and hence the computed value of $\zeta_{j,\,j+k}$ is a good estimate of $E(\boldsymbol{y}_j\,|\,\boldsymbol{x})^2$ as long as

$$E(\boldsymbol{y}_j\,|\,\boldsymbol{x}) \;\gg\; \kappa(\boldsymbol{x})^{1/2}\,P(n,k)\,Q_j(\boldsymbol{x})\mathbf{u}. \tag{12.22}$$

It is easy to see that $Q_j(\boldsymbol{x})$ is a sharp upper bound for $E(\boldsymbol{y}_0\,|\,\boldsymbol{x})$, so it follows that the computed estimate $\zeta_{j,\,j+k}$ is reliable until $E(\boldsymbol{y}_j\,|\,\boldsymbol{x})$ is of the order of $E(\boldsymbol{y}_0\,|\,\boldsymbol{x})\mathbf{u}$. Therefore, under assumptions (12.17), the inner stopping criterion is numerically stable as long as the contraction number γ is not too small (so j does not become too large), which is also recommended to avoid increasing the running time unnecessarily. In particular, since $n + m(\boldsymbol{x})\,n^{1/2} \leqslant n(1 + n^{1/2}) \approx n^{3/2}$, the inner stopping criterion works well in any basin $\mathcal{B}$ such that

$$\max\left(\frac{\sup_{\mathcal{B}}\kappa}{n^{1/2}},\, \sup_{\mathcal{B}}\kappa^\dagger\right) \;\ll\; \frac{1}{n^2\mathbf{u}}. \tag{12.23}$$

Algorithm 2 PCG-QMM.

INPUT: starting point $\boldsymbol{x}_0$, contraction number γ, and PCG stopping delay k.
OUTPUT: approximate solution to the optimization problem $\min_{\boldsymbol{x}} f(\boldsymbol{x})$.

for $p = 0, 1, 2, \ldots$ **do**
  $\boldsymbol{y}_0 = \boldsymbol{x}_p$
  $\boldsymbol{r}_0 = -\boldsymbol{w}(\boldsymbol{x}_p)$
  $\boldsymbol{p}_0 = \boldsymbol{s}_0 = \boldsymbol{M}(\boldsymbol{x}_p)^{-1}\boldsymbol{r}_0$
  $\beta_0 = \boldsymbol{r}_0^T \boldsymbol{s}_0$
  $\Delta_0 = \delta_0 = 0$
  for $j = 0, 1, 2, \ldots$ **do** ▷ PCG loop
    $\boldsymbol{q}_j = \boldsymbol{W}(\boldsymbol{x}_p)\boldsymbol{p}_j$
    $\alpha_j = \beta_j / (\boldsymbol{p}_j^T \boldsymbol{q}_j)$
    $\boldsymbol{y}_{j+1} = \boldsymbol{y}_j + \alpha_j \boldsymbol{p}_j$
    $\Delta_{j+1} = \Delta_j + \alpha_j \beta_j$
    $\delta_{j+1} = \delta_j + \alpha_j \beta_j$
    if $j \geqslant k$ **then**
      $\delta_{j+1} = \delta_{j+1} - \alpha_{j-k}\beta_{j-k}$
      if $\delta_{j+1} \leqslant \gamma \Delta_{j+1}$ **then** ▷ PCG stopping test
        $\boldsymbol{x}_{p+1} = \boldsymbol{y}_{j+1}$
        break ▷ exits the PCG loop
      end if
    end if
    $\boldsymbol{r}_{j+1} = \boldsymbol{r}_j - \alpha_j \boldsymbol{q}_j$
    $\boldsymbol{s}_{j+1} = \boldsymbol{M}(\boldsymbol{x}_p)^{-1}\boldsymbol{r}_{j+1}$
    $\beta_{j+1} = \boldsymbol{r}_{j+1}^T \boldsymbol{s}_{j+1}$
    $\boldsymbol{p}_{j+1} = \boldsymbol{s}_{j+1} + (\beta_{j+1}/\beta_j)\boldsymbol{p}_j$
  end for
end for

12.4. **Pseudocode.** Algorithm 2 generates the QMM recurrence $\boldsymbol{x}_{p+1} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x}_p)$ using a PCG solver that computes an approximate solution to the system $\boldsymbol{W}(\boldsymbol{x}_p)\boldsymbol{y} = \boldsymbol{v}(\boldsymbol{x}_p)$ starting from $\boldsymbol{x}_p$, so the initial residual is $\boldsymbol{r}_0 = \boldsymbol{v}(\boldsymbol{x}_p) - \boldsymbol{W}(\boldsymbol{x}_p)\boldsymbol{x}_p = -\boldsymbol{w}(\boldsymbol{x}_p)$, which equals $-\nabla f(\boldsymbol{x}_p)$ if f is differentiable at $\boldsymbol{x}_p$ (Proposition 3.2). We call this algorithm the PCG-QMM algorithm. It consists of two nested loops: the outer QMM iteration and the inner PCG iteration.

The implementation of the PCG stopping criterion described in Section 12.2 uses the local variables $\Delta_j := \zeta_{0,j}$ and $\delta_j := \zeta_{j-k,j}$. The stopping test at the $(j+1)$th iteration is obtained by replacing j by $j + 1 - k$ in (12.11):

$$\delta_{j+1} \leqslant \gamma \Delta_{j+1} \iff \sum_{i=j+1-k}^{j} \alpha_i \beta_i \leqslant \gamma \sum_{i=0}^{j} \alpha_i \beta_i. \tag{12.24}$$

If this inequality holds, then $\boldsymbol{y}_{j+1-k} \in \boldsymbol{\Phi}_{\gamma'}(\boldsymbol{x}_p)$ for some $\gamma' \approx \gamma$. Hence, since $E(\boldsymbol{y}_j | \boldsymbol{x}_p)$ decreases with j, the solution $\boldsymbol{y}_{j+1}$ returned by the PCG solver will usually belong to $\boldsymbol{\Phi}_{\gamma''}(\boldsymbol{x}_p)$ for some $\gamma'' < \gamma$.

12.4.1. *Least squares variant.* If the weighting matrices are normal equations matrices, that is, if they are given in a factored form $\boldsymbol{W}(\boldsymbol{x}) = \boldsymbol{A}(\boldsymbol{x})^T\boldsymbol{A}(\boldsymbol{x})$, then we can use PCG for least squares (PCGLS) to avoid explicitly computing $\boldsymbol{W}(\boldsymbol{x})$. It suffices to replace

line 8 by $\boldsymbol{q}_j := \boldsymbol{A}(\boldsymbol{x}_p)\boldsymbol{p}_j$,
line 9 by $\alpha_j := \beta_j / \|\boldsymbol{q}_j\|^2$,
line 20 by $\boldsymbol{r}_{j+1} := \boldsymbol{r}_j - \alpha_j \boldsymbol{A}(\boldsymbol{x}_p)^T\boldsymbol{q}_j$.

We call this variant the PCGLS-QMM algorithm.

12.4.2. *Outer termination test.* When f is $\mathscr{C}^1$, the convergence results in Sections 6–8 suggest to terminate the QMM iteration by monitoring the norm of the gradient. In our experience, a reliable criterion is

$$\|\nabla f(\boldsymbol{x}_p)\| \leqslant \epsilon \max(1, |f(\boldsymbol{x}_p)|), \tag{12.25}$$

where ϵ is a given tolerance. The term $\max(1, |f(\boldsymbol{x}_p)|)$ serves two purposes: to prevent the right-hand side from becoming too small if $f(\boldsymbol{x}_p)$ gets close to zero; and to ensure scale independence (the inequality $\|\nabla f\| \leqslant \epsilon|f|$ is unchanged if f is replaced by cf for any $c > 0$, as opposed to $\|\nabla f\| \leqslant \epsilon$).

12.4.3. *Continuation.* When f is nonconvex and/or approximates a nondifferentiable objective, it may be advantageous to guide the first QMM iterations by gradually increasing the optimization difficulty. This technique, which we call *continuation*, is reminiscent of graduated nonconvexity [42]. The design of a continuation scheme involves two stages:

(i) the construction of a finite sequence $(f_p)_{0 \leqslant p \leqslant q}$ of $\mathscr{C}^1$ *relaxed* objectives such that f_0 is convex, $f_q = f$, and the difficulty of minimizing f_p increases with p;
(ii) the construction of weighting functions $\boldsymbol{W}_p$ such that

$$f_p(\boldsymbol{y}|\boldsymbol{x}) := f_p(\boldsymbol{x}) + (\boldsymbol{y} - \boldsymbol{x})^T \nabla f_p(\boldsymbol{x}) + \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2_{\boldsymbol{W}_p(\boldsymbol{x})} \tag{12.26}$$

defines a surrogate for f_p.

The implementation is straightforward: in the first q outer iterations, $\boldsymbol{w}$ and $\boldsymbol{W}$ are replaced by ∇f_p and $\boldsymbol{W}_p$, respectively, and $\boldsymbol{M}(\boldsymbol{x}_p)$ is a preconditioner for the inner system matrix $\boldsymbol{W}_p(\boldsymbol{x}_p)$. (For the least squares variant, $\boldsymbol{W}_p(\boldsymbol{x})$ must be factored as $\boldsymbol{A}_p(\boldsymbol{x})^T\boldsymbol{A}_p(\boldsymbol{x})$.)

Continuation leads to deeper minima and/or accelerates convergence when the relaxed objectives f_p approximate f with increasing accuracy and the relaxed surrogates $f_p(\cdot|\cdot)$ are sufficiently close to the f_p.

13. Example applications

This section focuses on the application of CG-based QMM to the general problems of multidimensional scaling and regularized linear inversion. Each description begins with the construction of the surrogate followed by a discussion on the convergence results, which will be illustrated in Section 14.

13.1. **Multidimensional scaling.** Multidimensional scaling (MDS) refers to a family of methods for mapping similarity or dissimilarity data on pairs of objects (say m) into distances between points in a low-dimensional space (usually $\mathbb{R}^2$ or $\mathbb{R}^3$) [9,43]. A common way to carry out MDS is to minimize the so-called *raw stress* function

$$F(\boldsymbol{X}) := \sum_{i,j=1}^{m} c_{ij}\big(d_{ij} - \big\|\boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)}\big\|\big)^2, \tag{13.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{m\times q}$ is the coordinate matrix of the m objects in q dimensions, $\boldsymbol{X}^{(i)} := \boldsymbol{X}(i,:)^T$ (the transpose of the ith row of $\boldsymbol{X}$) contains the coordinates of object i, the weights c_{ij} are nonnegative, and d_{ij} (also nonnegative) is the dissimilarity between objects i and j.

13.1.1. *A surrogate for the stress function.* Up to an additive constant, $F(\boldsymbol{X})$ has the form

$$\sum_{i<j} c'_{ij}\big(d'_{ij} - \big\|\boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)}\big\|\big)^2 \tag{13.2}$$

with $c'_{ij} := c_{ij} + c_{ji}$ and $d'_{ij} := (c_{ij}d_{ij} + c_{ji}d_{ji})\mathrm{inv}(c'_{ij})$, where inv is the extended reciprocal function (8.9). Thus we can assume, without loss of generality, that the matrices $[c_{ij}]$ and $[d_{ij}]$ are symmetric and have zero diagonal. We also assume that

$$\sum_{i,j=1}^{m} c_{ij}d_{ij}^2 \;=\; 1 \tag{13.3}$$

(so $F(\boldsymbol{0}) = 1$) and that $[c_{ij}]$ is irreducible (so the original MDS problem cannot be decomposed into independent MDS subproblems).

Define

$$b_{ij}(\boldsymbol{X}) := c_{ij}d_{ij}\,\mathrm{inv}\big(\big\|\boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)}\big\|\big), \tag{13.4}$$

$$\langle \boldsymbol{X},\boldsymbol{Y}\rangle_{ij} := \big(\boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)}\big)^T\big(\boldsymbol{Y}^{(i)} - \boldsymbol{Y}^{(j)}\big). \tag{13.5}$$

By the Cauchy-Schwarz inequality,

$$b_{ij}(\boldsymbol{X})\langle \boldsymbol{X},\boldsymbol{Y}\rangle_{ij} \;\leqslant\; c_{ij}d_{ij}\big\|\boldsymbol{Y}^{(i)} - \boldsymbol{Y}^{(j)}\big\| \tag{13.6}$$

with equality if $\boldsymbol{Y} = \boldsymbol{X}$. Thus

$$\begin{aligned} F(\boldsymbol{Y}|\boldsymbol{X}) &:= 1 + \sum_{i,j=1}^{m}\big(-2b_{ij}(\boldsymbol{X})\langle \boldsymbol{X},\boldsymbol{Y}\rangle_{ij} + c_{ij}\big\|\boldsymbol{Y}^{(i)} - \boldsymbol{Y}^{(j)}\big\|^2\big) \\ &\geqslant F(\boldsymbol{Y}) \end{aligned} \tag{13.7}$$

and $F(\boldsymbol{X}|\boldsymbol{X}) = F(\boldsymbol{X})$. In other words, the quadratic function $F(\cdot|\boldsymbol{X})$ satisfies the domination and tangency conditions for a surrogate (see Definition 3.1). However, because of translational invariance, it is not positive definite. Before addressing this issue, we write $F(\cdot|\boldsymbol{X})$ in matrix form using the trace function. Recall that the Laplacian of an $m \times m$ symmetric nonnegative matrix $\boldsymbol{A} := [a_{ij}]$ is defined by

$$\boldsymbol{L}(\boldsymbol{A}) := \mathrm{diag}\left(\sum_{j=1}^{m} a_{ij}\right) - \boldsymbol{A}. \tag{13.8}$$

Clearly, $\boldsymbol{L}(\boldsymbol{A})$ is symmetric and diagonally dominant and has nonnegative diagonal entries. Hence, by the Gershgorin disc theorem, $\boldsymbol{L}(\boldsymbol{A})$ is positive semidefinite. Let

$\mathrm{tr}(\cdot)$ be the trace function and let $\boldsymbol{e}_{ij} := \boldsymbol{e}_i - \boldsymbol{e}_j$, where $\boldsymbol{e}_i$ denotes the ith standard basis vector in $\mathbb{R}^m$. It is easy to see that

$$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle_{ij} \;=\; \mathrm{tr}\big(\boldsymbol{Y}^T \boldsymbol{e}_{ij} \boldsymbol{e}_{ij}^T \boldsymbol{X}\big) \quad \text{and} \quad 2\boldsymbol{L}(\boldsymbol{A}) \;=\; \sum_{i,j=1}^{m} a_{ij} \boldsymbol{e}_{ij} \boldsymbol{e}_{ij}^T, \tag{13.9}$$

from which it follows that

$$\sum_{i,j=1}^{m} a_{ij} \langle \boldsymbol{X}, \boldsymbol{Y} \rangle_{ij} \;=\; 2\mathrm{tr}(\boldsymbol{Y}^T \boldsymbol{L}(\boldsymbol{A}) \boldsymbol{X}). \tag{13.10}$$

Thus, with $\boldsymbol{B}(\boldsymbol{X}) := [b_{ij}(\boldsymbol{X})]$ and $\boldsymbol{C} := [c_{ij}]$, we have

$$F(\boldsymbol{Y} | \boldsymbol{X}) \;=\; 1 - 4\mathrm{tr}(\boldsymbol{Y}^T \boldsymbol{L}(\boldsymbol{B}(\boldsymbol{X})) \boldsymbol{X}) + 2\mathrm{tr}(\boldsymbol{Y}^T \boldsymbol{L}(\boldsymbol{C}) \boldsymbol{Y}). \tag{13.11}$$

We remove the translational degree of freedom by setting, say, $\boldsymbol{X}_m = \boldsymbol{0}$. The MDS problem is then to minimize the stress

$$F' := F \circ \boldsymbol{\mu}, \qquad \boldsymbol{\mu} : \boldsymbol{X}' \in \mathbb{R}^{(m-1)\times q} \longmapsto \begin{bmatrix} \boldsymbol{X}' \\ 0 \cdots 0 \end{bmatrix} \in \mathbb{R}^{m \times q}. \tag{13.12}$$

A surrogate for F' is

$$\begin{aligned} F'(\boldsymbol{Y}' | \boldsymbol{X}') \;&:=\; F(\boldsymbol{\mu}(\boldsymbol{Y}') \,|\, \boldsymbol{\mu}(\boldsymbol{X}')) \\ &= 1 - 4\mathrm{tr}(\boldsymbol{Y}'^T \boldsymbol{U}(\boldsymbol{X}') \boldsymbol{X}') + 2\mathrm{tr}(\boldsymbol{Y}'^T \boldsymbol{V} \boldsymbol{Y}'), \end{aligned} \tag{13.13}$$

where $\boldsymbol{U}(\boldsymbol{X}')$ and $\boldsymbol{V}$ are the $(m-1) \times (m-1)$ leading principal submatrices (that is, the submatrices obtained by deleting the last row and the last column) of $\boldsymbol{L}(\boldsymbol{B}(\boldsymbol{\mu}(\boldsymbol{X}')))$ and $\boldsymbol{L}(\boldsymbol{C})$, respectively. It remains to explain why $\boldsymbol{V}$ is positive definite. Since $\boldsymbol{C}$ is irreducible, it follows from the matrix-tree theorem (see, e.g., [44, Section II.3]) that all the cofactors of $\boldsymbol{L}(\boldsymbol{C})$ are equal and positive. So all the $(m-1) \times (m-1)$ submatrices of $\boldsymbol{L}(\boldsymbol{C})$ are nonsingular. Furthermore, since $\boldsymbol{L}(\boldsymbol{C})$ is positive semidefinite, all its principal submatrices are positive semi-definite. Therefore the $(m-1) \times (m-1)$ principal submatrices of $\boldsymbol{L}(\boldsymbol{C})$ are positive definite.

13.1.2. *Multidimensional scaling by PCG-QMM.* Let $n := (m-1)q$ and let $\boldsymbol{\nu}$ be the columnwise vectorization map from $\mathbb{R}^{(m-1)\times q}$ to $\mathbb{R}^n$, so $\boldsymbol{x} \in \mathbb{R}^n$ can be identified with $\boldsymbol{X}' := \boldsymbol{\nu}^{-1}(\boldsymbol{x})$ and hence with

$$\boldsymbol{X} := \boldsymbol{\mu}(\boldsymbol{\nu}^{-1}(\boldsymbol{x})). \tag{13.14}$$

Minimizing the stress F' in (13.12) using the surrogate (13.13) is equivalent to minimizing

$$f(\boldsymbol{x}) \;:=\; F\left(\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_m & \cdots & \boldsymbol{x}_{(q-1)m-q+2} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{x}_{m-1} & \boldsymbol{x}_{2(m-1)} & \cdots & \boldsymbol{x}_{q(m-1)} \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \end{bmatrix} \right) \tag{13.15}$$

using

$$f(\boldsymbol{y} | \boldsymbol{x}) \;:=\; 1 - 4\boldsymbol{y}^T (\boldsymbol{I}_q \otimes \boldsymbol{U}(\boldsymbol{x})) \boldsymbol{x} + 2\boldsymbol{y}^T (\boldsymbol{I}_q \otimes \boldsymbol{V}) \boldsymbol{y}, \tag{13.16}$$

where $\boldsymbol{I}_q$ is the $q \times q$ identity matrix and $\otimes$ is the Kronecker product operator. The above surrogate can be expressed in the canonical form (3.1) with weighting functions given by

$$\boldsymbol{w}(\boldsymbol{x}) \;=\; 4(\boldsymbol{I}_q \otimes (\boldsymbol{V} - \boldsymbol{U}(\boldsymbol{x})))\boldsymbol{x} \quad \text{and} \quad \boldsymbol{W}(\boldsymbol{x}) \;=\; 4\boldsymbol{I}_q \otimes \boldsymbol{V}. \tag{13.17}$$

Thus the inner system (12.1) is

$$(\boldsymbol{I}_q \otimes \boldsymbol{V})\boldsymbol{y} \,=\, (\boldsymbol{I}_q \otimes \boldsymbol{U}(\boldsymbol{x}))\boldsymbol{x}. \tag{13.18}$$

It is important to emphasize that the coefficient matrix $\boldsymbol{I}_q \otimes \boldsymbol{V}$ does not depend on $\boldsymbol{x}$, so we do not have to worry about the cost of computing the preconditioner.

The objective f is semialgebraic and hence tame subanalytic. It is analytic on the open set $\mathcal{O} \subset \mathbb{R}^n$ consisting of all points $\boldsymbol{x}$ such that $\boldsymbol{X}_i \neq \boldsymbol{X}_j$ whenever $c_{ij}\, d_{ij} > 0$, but it is nondifferentiable at any point $\boldsymbol{x} \notin \mathcal{O}$. The local minimizers of f, however, all lie in $\mathcal{O}$ [45]; so we can ignore the points of nondifferentiability, as we now argue. First, note that the surrogate function $f(\cdot\,|\boldsymbol{x})$ is well-defined for all $\boldsymbol{x}$, so the PCG-QMM algorithm does not fail if $\boldsymbol{x}_p \notin \mathcal{O}$ for some p. Since one of the objects is fixed at the origin, it follows from the irreducibility of $\boldsymbol{C}$ that f is coercive. Thus the algorithm generates bounded sequences, and the convergence results in Section 9 hold provided that $\liminf_p \operatorname{dist}(\boldsymbol{x}_p, \mathbb{R}^n \setminus \mathcal{O}) > 0$. This is the case in practice, as the monotone convergence of $\{f(\boldsymbol{x}_p)\}$ implies that points that are not minimizers are not stable against round-off errors. In fact, no point of nondifferentiability was encountered in our experiments. Should this occur, we can ensure the generation of QMM sequences in $\mathcal{O}$ via a slight modification: if, starting from $\boldsymbol{x}_p$, the PCG solver stops at a point $\boldsymbol{y}_{j+1} \notin \mathcal{O}$, then set $\boldsymbol{x}_{p+1}$ to be the first (or any) PCG iterate $\boldsymbol{y}_{j+1} \in \mathcal{O}$ (such exists because $\mathbb{R}^n \setminus \mathcal{O}$ is a finite union of subspaces of dimension $n-q$ and the conjugate directions have no connection with the directions of these subspaces).

Regarding the convergence rate, Corollary 10.3 does not apply because no stationary point is isolated (since f is invariant under linear isometries of the coordinate space). We can only assert that the convergence rate is either R-linear or -sublinear (Corollary 9.5), and that it is R-linear when there is a minimum angle between $\boldsymbol{x}_p - \boldsymbol{x}$ and the tangent space of $\mathcal{S}_f$ at $\boldsymbol{x}$ (Theorem 10.2 together with the fact that $\nabla^2 f$ does not vanish on $\mathcal{O}$ [46]).

13.1.3. *Comparison with the SMACOF algorithm.* The semidefinite majorizing function $F(\cdot\,|\boldsymbol{X})$ in (13.7) goes back to [47] and is the cornerstone of the well-known SMACOF algorithm [9, 48, 49], which stands for "Scaling by MAjorizing a COmplicated Function."

Let $\Phi^+ \colon \mathbb{R}^{m\times q} \to \mathbb{R}^{m\times q}$ be defined by

$$\Phi^+(\boldsymbol{X}) \,:=\, \boldsymbol{L}(\boldsymbol{C})^+ \boldsymbol{L}(\boldsymbol{B}(\boldsymbol{X}))\boldsymbol{X}, \tag{13.19}$$

where $\boldsymbol{L}(\boldsymbol{C})^+$ is the Moore-Penrose generalized inverse of $\boldsymbol{L}(\boldsymbol{C})$. The configuration $\Phi^+(\boldsymbol{X})$ is the minimizer of $F(\cdot\,|\boldsymbol{X})$ with minimum Frobenius norm, and the SMACOF algorithm consists of the iteration $\boldsymbol{X}_{p+1} = \Phi^+(\boldsymbol{X}_p)$. With the proviso of no rounding errors, the convergence properties of SMACOF sequences are limited to the following [48]:

(i) $\mathcal{L}_{\{\boldsymbol{X}_p\}}$ (the limit set of $\{\boldsymbol{X}_p\}$) is a flat continuum, and $\{F(\boldsymbol{X}_p)\}$ decreases to the value of F on $\mathcal{L}_{\{\boldsymbol{X}_p\}}$.

(ii) Any point in $\mathcal{L}_{\{\boldsymbol{X}_p\}}$ at which F is differentiable is a stationary point of F.

(iii) $\lim_p \|\boldsymbol{L}(\boldsymbol{C})^{1/2}(\boldsymbol{X}_{p+1} - \boldsymbol{X}_p)\| = 0$.

These properties are similar to the subconvergence properties in Theorem 5.2 (except that $\{\boldsymbol{X}_{p+1} - \boldsymbol{X}_p\}$ converges to $\boldsymbol{0}$ with respect to a seminorm). So PCG-QMM offers stronger convergence guarantees than does SMACOF, with the further advantage of numerical stability.

13.2. **Regularized linear inversion.** Linear inverse problems refer to the recovery of a signal $\boldsymbol{x}^\star \in \mathbb{R}^n$ (typically a temporal sequence, an image, or a discretized volume) from data of the form

$$\boldsymbol{d} := \boldsymbol{D}\boldsymbol{x}^\star + \boldsymbol{\varepsilon}, \tag{13.20}$$

where $\boldsymbol{D} \in \mathbb{R}^{m\times n}$ models the deterministic part of the measurement process and $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ is a noise vector whose components are realizations of independent zero-mean random variables. This observation model covers many imaging applications [50–52], the most prominent being deconvolution and tomographic reconstruction. A common approach to estimating the original signal $\boldsymbol{x}^\star$ is to minimize an objective of the form (see, e.g., [10, 53])

$$f(\boldsymbol{x}) := \sum_{i=1}^{m} \theta_i\big(\big|[\boldsymbol{D}\boldsymbol{x} - \boldsymbol{d}]_i\big|\big) + \sum_{i=m+1}^{M} \theta_i\big(\|\boldsymbol{A}_i\boldsymbol{x}\|\big), \tag{13.21}$$

where $\theta_1, \dots, \theta_M$ are increasing functions on $\mathbb{R}_+$, $[\,\cdot\,]_i$ is the ith coordinate projection, and the $\boldsymbol{A}_i$'s are matrices with possibly different ranges. The first sum—the *fidelity* term—favors solutions consistent with the data. Usually, $\theta_1 = \cdots = \theta_m = \vartheta$ with $\vartheta(t) \propto t^2$ or with $\vartheta(t)$ convex and asymptotically linear (the former choice yields the squared ℓ_2-norm of the residual, while the latter reduces the sensitivity to outliers). The second sum—the *regularization* term—incorporates prior knowledge about the original signal. Often, $\{\boldsymbol{A}_i\}_{m<i\leqslant M}$ is a discrete gradient operator; more sophisticated alternatives include higher-order differential operators [54] and sparsifying transforms such as wavelets [55], framelets [56], and patch dictionaries [57].

13.2.1. *A surrogate for the inversion objective.* The objective (13.21) has the general form

$$f(\boldsymbol{x}) := \sum_{i=1}^{M} \theta_i\big(\|\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{a}_i\|\big) \tag{13.22}$$

with $\boldsymbol{A}_i \in \mathbb{R}^{n_i\times n}$ and $\boldsymbol{a}_i \in \mathbb{R}^{n_i}$ for some positive integer n_i. We call this function the *inversion* objective and we call the θ_i's the *potentials* (a term borrowed from the Bayesian interpretation of regularization [58]). Restricting ourselves to the $\mathscr{C}^1$ case, the minimal assumptions on the potentials for constructing the surrogate are as follows [53]:

(i) $\theta_1, \dots, \theta_M$ are increasing and $\mathscr{C}^1$;

(ii) the functions $\theta_i^\dagger : t \in (0, +\infty) \longmapsto \theta_i'(t)/t$ are decreasing and bounded.

Such potentials are called *admissible*. Apart from the square function, one usually distinguishes three categories of admissible potentials: (i) convex and asymptotically linear, (ii) nonconvex and unbounded, and (iii) bounded (common examples are listed in [10, Appendix A]). We recall that admissible potentials are twice right-differentiable at zero with $\theta_i'(0) = 0$ and $\theta_i''(0) > 0$, and hence behave quadratically near zero.

We assume from now on that the potentials θ_i are all admissible. We then say that the inversion objective f is admissible. Let

$$\boldsymbol{A} := \begin{bmatrix} \boldsymbol{A}_1 \\ \vdots \\ \boldsymbol{A}_M \end{bmatrix} \in \mathbb{R}^{N\times n} \quad \text{and} \quad \boldsymbol{a} := \begin{bmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_M \end{bmatrix} \in \mathbb{R}^{N}, \qquad N := \sum_{i=1}^{M} n_i, \tag{13.23}$$

and let $\boldsymbol{Q}(\boldsymbol{x})$ be the $N \times N$ nonnegative diagonal matrix defined by

$$\boldsymbol{Q}(\boldsymbol{x}) := \operatorname{diag}(q_i(\boldsymbol{x})\boldsymbol{I}_{n_i}), \qquad q_i(\boldsymbol{x}) := \theta_i^{\dagger}\big(\|\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{a}_i\|\big), \tag{13.24}$$

with the convention that $\theta_i^{\dagger}(0) = \lim_{t\to 0^+} \theta_i^{\dagger}(t)$. The gradient of f is given by

$$\nabla f(\boldsymbol{x}) \;=\; \boldsymbol{A}^T\boldsymbol{Q}(\boldsymbol{x})(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{a}). \tag{13.25}$$

Define

$$f(\boldsymbol{y}\,|\,\boldsymbol{x}) \;:=\; f(\boldsymbol{x}) + (\boldsymbol{y}-\boldsymbol{x})^T\nabla f(\boldsymbol{x}) + \tfrac{1}{2}\|\,\boldsymbol{y}-\boldsymbol{x}\,\|^2_{\boldsymbol{W}(\boldsymbol{x})}, \tag{13.26a}$$

$$\boldsymbol{W}(\boldsymbol{x}) \;:=\; \boldsymbol{A}^T\boldsymbol{Q}(\boldsymbol{x})\boldsymbol{A}. \tag{13.26b}$$

In [53, Proposition 2.4] it is shown that the weighting matrix $\boldsymbol{W}(\boldsymbol{x})$ is positive definite for all $\boldsymbol{x}$ (so $f(\cdot\,|\,\cdot)$ is a surrogate for f) if and only if

$$\bigcap_{\substack{i=1 \\ \theta_i \text{ strictly increasing}}}^{M} \operatorname{null}(\boldsymbol{A}_i) \;=\; \{\boldsymbol{0}\}. \tag{13.27}$$

This condition is satisfied if f is coercive; regardless, its violation can be considered a failure of the regularization scheme.

13.2.2. *Regularized inversion by PCGLS-QMM.* The weighting matrix can be factored as

$$\boldsymbol{W}(\boldsymbol{x}) \;=\; \boldsymbol{A}(\boldsymbol{x})^T\boldsymbol{A}(\boldsymbol{x}), \qquad \boldsymbol{A}(\boldsymbol{x}) := \boldsymbol{Q}(\boldsymbol{x})^{1/2}\boldsymbol{A}, \tag{13.28}$$

so we can use the least squares variant of the PCG-QMM algorithm (see Section 12.4.1). The behavior of the PCGLS-QMM algorithm depends on the properties of the potentials. Two special cases deserve attention: when the θ_i's are piecewise analytic, as are the potentials encountered in the literature, and when they are convex.

Definition 13.1. A potential θ is said to be *piecewise analytic* if there is a finite partition of $\mathbb{R}_+$ into intervals $\mathcal{I}_1, \ldots, \mathcal{I}_l$ such that for each $j \in \{1, \ldots, l\}$ the restriction $\theta|_{\mathcal{I}_j}$ has an analytic extension on a neighborhood of the closure of $\mathcal{I}_j$ (that is, there exist an open interval $\mathcal{O}_j \supset \overline{\mathcal{I}_j}$ and an analytic function $g_j : \mathcal{O}_j \to \mathbb{R}$ such that $\theta(t) = g_j(t)$ for all $t \in \mathcal{I}_j$).

Theorem 13.2. *Let $\{\boldsymbol{x}_p\}$ be a bounded $\mathscr{C}^0$-QMM sequence with admissible inversion objective f, and assume that the potentials are piecewise analytic. Then $\{\boldsymbol{x}_p\}$ converges at least R-sublinearly to a stationary point of f.*

Proof. By Theorem 6.8 in [10] the objective is tame subanalytic. Therefore $\{\boldsymbol{x}_p\}$ converges to a stationary point (Theorem 8.4) and the convergence rate is R-linear or -sublinear (Corollary 9.5). □

Theorem 13.3. *Let $\{\boldsymbol{x}_p\}$ be a $\mathscr{C}^0$-QMM sequence with admissible inversion objective f, and assume that the potentials are convex. Consider the following conditions:*

$$\bigcap_{\substack{i=1 \\ \lim_{t\to+\infty}\theta_i(t)=+\infty}}^{M} \operatorname{null}(\boldsymbol{A}_i) \;=\; \{\boldsymbol{0}\}, \tag{13.29}$$

$$\bigcap_{\substack{i=1 \\ \theta_i \text{ strictly convex}}}^{M} \operatorname{null}(\boldsymbol{A}_i) \;=\; \{\boldsymbol{0}\}. \tag{13.30}$$

(i) *If* (13.29) *holds, then $\{\boldsymbol{x}_p\}$ converges to* $\arg\min f$.
(ii) *If* (13.30) *holds, then $\{\boldsymbol{x}_p\}$ converges to the global minimizer of f.*
(iii) *If, in addition to* (13.30)*, the potentials are $\mathscr{C}^2$, then the convergence rate is R-linear.*

Proof. (i) Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct points in $\mathbb{R}^n$ and let $\alpha \in (0,1)$. Since the potentials are increasing and convex, we have for all i that

$$\begin{aligned} &\theta_i\big(\|\boldsymbol{A}_i(\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y})-\boldsymbol{a}_i\|\big) \\ &\qquad\leqslant \alpha\theta_i\big(\|\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{a}_i\|\big)+(1-\alpha)\theta_i\big(\|\boldsymbol{A}_i\boldsymbol{y}-\boldsymbol{a}_i\|\big), \end{aligned} \tag{13.31}$$

and hence

$$f(\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}) \;\leqslant\; \alpha f(\boldsymbol{x})+(1-\alpha)f(\boldsymbol{y}).$$

So f is convex. Furthermore, (13.29) is equivalent to f being coercive [53, Proposition 2.5]. The result then follows from Corollary 6.3(ii).

(ii) Since (13.30) implies (13.29), f is coercive and $\{\boldsymbol{x}_p\}$ converges to $\arg\min f$. So it suffices to show that f is strictly convex. There is an index i such that θ_i is strictly convex and $\boldsymbol{A}_i(\boldsymbol{y}-\boldsymbol{x}) \neq \boldsymbol{0}$. Let

$$\boldsymbol{v}_i \;:=\; \boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{a}_i \quad\text{and}\quad \boldsymbol{w}_i \;:=\; \boldsymbol{A}_i\boldsymbol{y}-\boldsymbol{a}_i.$$

If $\|\boldsymbol{v}_i\| \neq \|\boldsymbol{w}_i\|$, then inequality (13.31) is strict. Now suppose that $\|\boldsymbol{v}_i\| = \|\boldsymbol{w}_i\|$. Then (13.31) becomes

$$\theta_i\big(\|\alpha\boldsymbol{v}_i+(1-\alpha)\boldsymbol{w}_i\|\big) \;\leqslant\; \theta_i\big(\|\boldsymbol{v}_i\|\big). \tag{13.32}$$

It is easy to check that

$$2\boldsymbol{v}_i^T(\boldsymbol{w}_i-\boldsymbol{v}_i) \;=\; -\|\boldsymbol{w}_i-\boldsymbol{v}_i\|^2.$$

Using this identity, we have

$$\begin{aligned} \|\alpha\boldsymbol{v}_i+(1-\alpha)\boldsymbol{w}_i\|^2 \;&=\; \|\boldsymbol{v}_i+(1-\alpha)(\boldsymbol{w}_i-\boldsymbol{v}_i)\|^2 \\ &=\; \|\boldsymbol{v}_i\|^2+2(1-\alpha)\boldsymbol{v}_i^T(\boldsymbol{w}_i-\boldsymbol{v}_i)+(1-\alpha)^2\|\boldsymbol{w}_i-\boldsymbol{v}_i\|^2 \\ &=\; \|\boldsymbol{v}_i\|^2-\alpha(1-\alpha)\|\boldsymbol{w}_i-\boldsymbol{v}_i\|^2 \\ &<\; \|\boldsymbol{v}_i\|^2. \end{aligned}$$

Since θ_i is strictly increasing, it follows that inequality (13.32) is strict.

(iii) It can be shown that if the θ_i's are $\mathscr{C}^2$, then f is $\mathscr{C}^2$ with Hessian

$$\nabla^2 f(\boldsymbol{x}) \;=\; \sum_{i=1}^{M} \theta_i''\big(\|\boldsymbol{A}_i\boldsymbol{x}-\boldsymbol{a}_i\|\big)\boldsymbol{A}_i^T\boldsymbol{A}_i.$$

Let $\boldsymbol{y} \neq \boldsymbol{0}$. There is an index i such that θ_i is strictly convex and $\boldsymbol{A}_i\boldsymbol{y} \neq \boldsymbol{0}$. Hence, for any $\boldsymbol{x}$, we have

$$\boldsymbol{y}^T \nabla^2 f(\boldsymbol{x}) \boldsymbol{y} \geqslant \theta_i''\big(\|\boldsymbol{A}_i\boldsymbol{x} - \boldsymbol{a}_i\|\big)\|\boldsymbol{A}_i\boldsymbol{y}\|^2 > 0.$$

Thus the Hessian is nowhere zero, and the result follows from Corollary 10.3. □

13.2.3. *Continuation.* The design of a continuation scheme (see Section 12.4.3) is facilitated by the form of the inversion objective:

(i) A sequence $(f_p)_{0 \leqslant p \leqslant q}$ of relaxed objectives is constructed by replacing some or all of the potentials by relaxed potentials $\theta_{i,p}$ such that $\theta_{i,q} = \theta_i$ and $\theta_{i,p}$ approximates θ_i with increasing accuracy.

(ii) The relaxed surrogates $f_p(\cdot|\cdot)$ are obtained by defining $\boldsymbol{Q}_p$ analogously to $\boldsymbol{Q}$ using the relaxed potentials, and substituting into (13.25) and (13.28) to get ∇f_p and $\boldsymbol{W}_p$.

Improvement in the quality of the solutions and acceleration of convergence are usually achieved with relaxed $\mathscr{C}^2$ potentials following two guidelines. First, if θ_i is nonconvex then the maximum concavity of $\theta_{i,p}$ (that is, the quantity $-\inf_{t \geqslant 0} \theta_{i,p}''(t)$) should increase to that of θ_i. In this way the nonconvexity of f increases with the number of iterations, and thus so does the optimization difficulty. Second, if θ_i approximates a potential whose first derivative does not vanish at zero (a case in point is when f approximates a nondifferentiable objective) then the sequence of second derivatives $(\theta_{i,p}''(0))_p$ should increase to $\theta_i''(0)$. This speeds up convergence by gradually reducing the range over which $\theta_{i,p}$ behaves quadratically.

14. Experiments

In this section we present experiments on graph layout and X-ray tomography as instances of multidimensional scaling and regularized inversion, respectively. The PCG- and PCGLS-QMM algorithms are implemented in MATLAB and run on a PC with an Intel Core i7-9850H CPU and 64 GB DDR4 2666 MHz RAM.

14.1. **Graph layout.** We consider the problem of producing aesthetically pleasing layouts of undirected graphs by mapping the edges to line segments in the plane (see [59, chapter 10] and [60] for an introduction to this subject). To do so, we seek to minimize the energy function introduced by Kamada and Kawai [61], which is a weighted sum of the squared differences of the Euclidean and graph-theoretic distances between vertices.

Let $\mathscr{G}$ be an undirected connected graph with vertex set $\mathcal{V} := \{v_1, \dots, v_m\}$ and edge set $\mathcal{E}$ (a set of unordered pairs of vertices). Let $\omega : \mathcal{E} \to (0, +\infty)$ be a weight function specifying the ideal edge lengths in the layout, and define the dissimilarity d_{ij} between the vertices v_i and v_j as the length of the shortest weighted path(s) between them:

$$d_{ij} := \min_{\boldsymbol{\pi} \text{ a } (v_i, v_j)\text{-path}} \sum_{e \in \boldsymbol{\pi}} \omega(e) \tag{14.1}$$

($d_{ij} > 0$ for all i, j, since $\mathscr{G}$ is connected). The Kamada-Kawai energy is a special case of the raw stress function (13.1) in which $q = 2$ and the matrix $[c_{ij}]$ is

TABLE 1. Graphs considered in the experiments: number of vertices (m), number of edges (card $\mathcal{E}$), and spectral condition number (κ) of the corresponding inner system matrix.

	m	card $\mathcal{E}$	κ
`1138_bus`	1138	1458	3.58×10^3
`tuma2`	12992	20925	2.58×10^4
`filter2D`	1668	4541	3.49×10^3
`rajat19`	819	1151	1.07×10^4

proportional to $[1/{d_{ij}}^2]$, that is,

$$F(\boldsymbol{X}) \propto \sum_{i,j=1}^{m} \Big(1 - \big\| \boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)} \big\| / d_{ij}\Big)^2, \tag{14.2}$$

where $\boldsymbol{X}^{(i)}$ and $\boldsymbol{X}^{(j)}$ contain the two-dimensional cartesian coordinates of v_i and v_j, respectively. The corresponding objective f and its surrogate are defined by (13.15) and (13.16).

We use the weight function proposed in [62], which assigns to each edge $\{v, v'\}$ the number of vertices that are neighbors of v or v', but not both:

$$\omega(\{v, v'\}) := \operatorname{card}\Big(\mathcal{N}(v) \cup \mathcal{N}(v') \setminus \mathcal{N}(v) \cap \mathcal{N}(v')\Big), \tag{14.3}$$

where $\mathcal{N}(v) := \{u \in \mathcal{V} : \{u, v\} \in \mathcal{E}\}$ (and similarly for v'). In other words, each edge is weighted by the cardinality of the symmetric difference of the neighborhoods of its endvertices. This penalizes dense aggregations around high-degree vertices, which produces more aesthetically pleasing layouts than does uniform weighting.

Figures 3 and 4 show examples of layouts produced by the PCG-QMM algorithm with a contraction number $\gamma = 10^{-6}$, a stopping delay $k = 5$, and an outer termination tolerance $\epsilon = 10^{-6}$ (see (12.25)). The adjacency matrices of these graphs are from the SuiteSparse Matrix Collection [63]. We will investigate the behavior of PCG-QMM with the graphs shown in Figure 3; Table 1 gives their numbers of vertices and edges together with the spectral condition number of their corresponding inner system matrix $\boldsymbol{I}_2 \otimes \boldsymbol{V}$ (see (13.18)). Each of these four graphs is assigned a fixed initial layout whose vertices are randomly distributed in $[-r, r]^2$, $r := \frac{1}{2} \max_{i,j} d_{ij}$, so the same starting point is used in all the experiments. The preconditioner $\boldsymbol{M} := \boldsymbol{H}\boldsymbol{H}^T$ is an incomplete Cholesky factorization of $\boldsymbol{I}_2 \otimes \boldsymbol{V}$ with a drop tolerance of 10^{-2} for `1138 bus` and `filter2D`, and 10^{-3} for `tuma2` and `rajat19`.

14.1.1. *Effects of the accuracy of the PCG solver.* Let N_{MM} denote the number of outer iterations to termination and let j_p denote the last PCG iteration number at the $(p{+}1)$th outer iteration (so $\boldsymbol{x}_{p+1} = \boldsymbol{y}_{j_p+1}$). The total number of PCG iterations is then

$$N_{\mathrm{CG}} := \sum_{p=0}^{N_{\mathrm{MM}}-1} (j_p + 1) \geqslant (k+1) N_{\mathrm{MM}}. \tag{14.4}$$

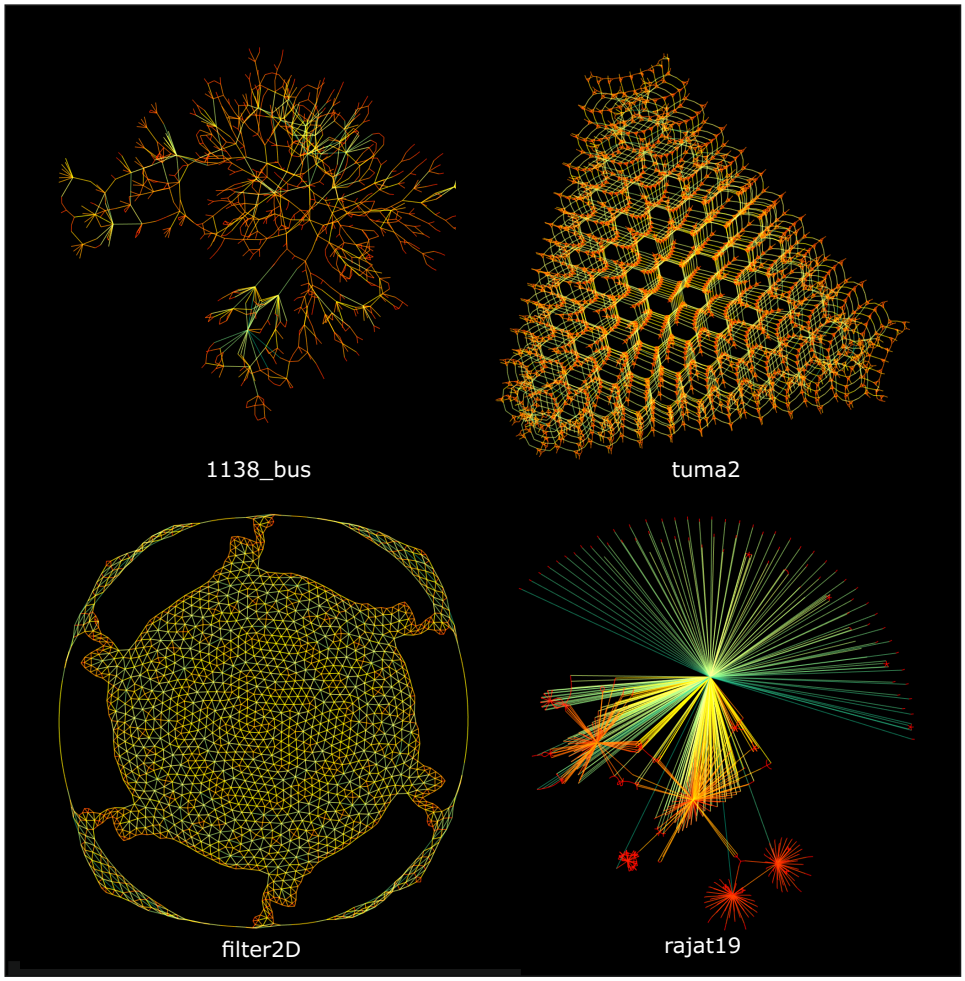

FIGURE 3. Examples of layouts produced by the PCG-QMM algorithm. The color scheme ranges from red (shortest edges) to yellow to green (longest edges).

Table 2 gives N_{MM}, $f(\boldsymbol{x}_{N_{\mathrm{MM}}})$ (the final objective value), $N_{\mathrm{CG}}/N_{\mathrm{MM}}$ (the mean number of PCG iterations per outer iteration), and the running time T for decreasing values of γ when $k = 5$ and $\epsilon = 10^{-6}$. As expected, N_{MM} and $f(\boldsymbol{x}_{N_{\mathrm{MM}}})$ stabilize as $\gamma \to 0$, while $N_{\mathrm{CG}}/N_{\mathrm{MM}}$ and T increase as γ decreases. Furthermore, the final layouts $\boldsymbol{X}_{N_{\mathrm{MM}}}$ are visually indistinguishable from those displayed in Figure 3, independently of γ.

Figure 5 plots N_{MM} and T as functions of a fixed number J of PCG iterations per outer iteration (so $N_{\mathrm{CG}} = N_{\mathrm{MM}} J$ and increasing J amounts to decreasing γ). Both N_{MM} and T grow rapidly when J decreases towards 1; thus the stopping delay is also a safeguard to limit the running time when γ is too large ($\gamma \geqslant 0.1$, say). By construction, the running time is nearly linear in N_{MM} and in N_{CG}:

$$T \approx c_1 N_{\mathrm{MM}} + c_2 N_{\mathrm{CG}}, \tag{14.5}$$

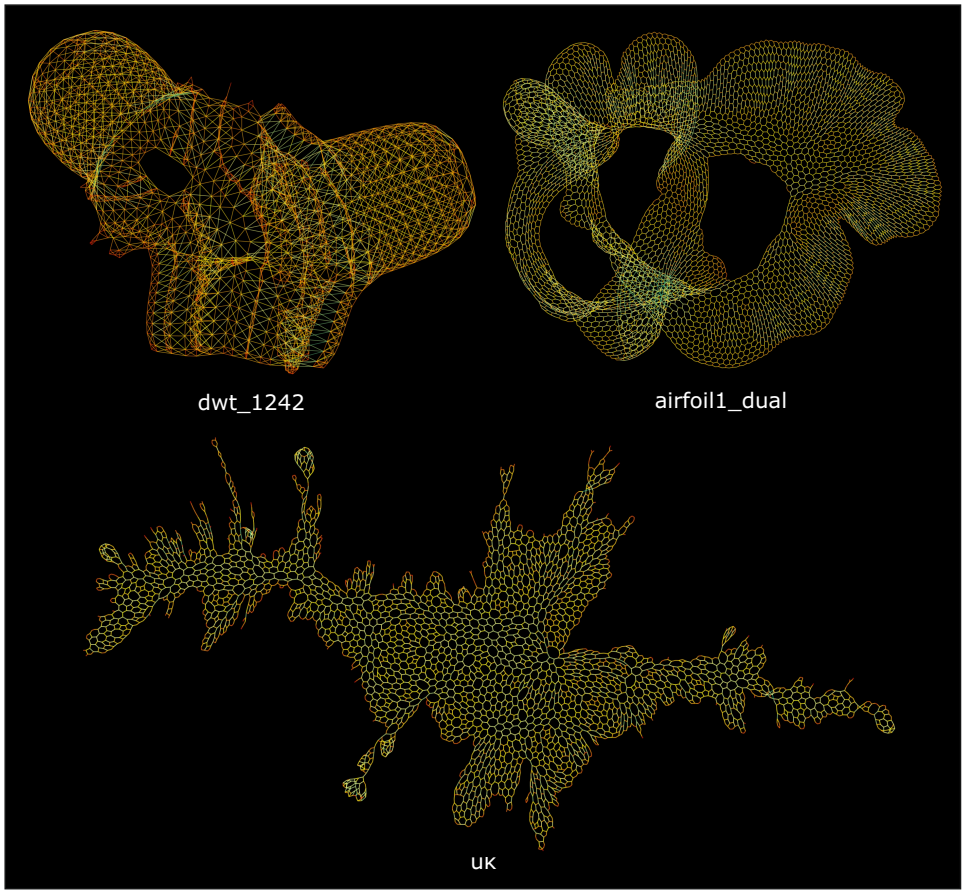

FIGURE 4. Examples of layouts produced by the PCG-QMM algorithm. The color scheme ranges from red (shortest edges) to yellow to green (longest edges).

where c_1 and c_2 are positive scalars depending on $\mathscr{G}$. This is illustrated by the solid curves in Figure 5(b), which represent the estimated running time $(c_1+c_2 J)N_{\mathrm{MM}}(J)$ for c_1 and c_2 obtained by linear regression. When J is large enough so that N_{MM} is approximately constant, T increases linearly with J and hence increases with decreasing γ. So overall there is a balance between the inner accuracy (controlled by k and γ) and the running time. In our experience, taking $k = 4$ or 5 and $\gamma \in [10^{-4}, 10^{-2}]$ is a good initial compromise.

14.1.2. *Behavior in the long run.* There is a limit to the accuracy of the gradient norm in finite precision arithmetic: $\|\nabla f(\boldsymbol{x}_p)\|$ eventually stagnates and we then say that the maximum accuracy has been attained. This is illustrated in Figure 6, which plots $\|\nabla f(\boldsymbol{x}_p)\|$ versus p for different values of the contraction number (the vertical dashed lines indicate the number N_{MM} of outer iterations to termination when $\epsilon = 10^{-6}$). The maximum accuracy is reached after about 4500 iterations for `1138 bus`, 600 iterations for `filter2D`, 6×10^5 iterations for `rajat19`, and in between 6000 and 11000 iterations for `tuma2`. In all cases, the mean value at

TABLE 2. Behavior of the PCG-QMM graph layout algorithm as a function of the contraction number (stopping delay $k = 5$, outer termination tolerance $\epsilon = 10^{-6}$).

γ	N_{MM}	$f(\boldsymbol{x}_{N_{\mathrm{MM}}})$	$N_{\mathrm{CG}}/N_{\mathrm{MM}}$	T
		`1138_bus`		
10^{-2}	304	42346.*7318029*	6.5	2.6s
10^{-4}	291	42346.1*555027*	10.6	2.9s
10^{-6}	278	42346.1637*648*	13.8	2.9s
10^{-8}	278	42346.163765	15.4	3.1s
10^{-10}	278	42346.1637655	16.7	3.2s
		`tuma2`		
10^{-2}	78	414738*8.34310*	7.8	1min 50s
10^{-4}	76	41473*97.29226*	14.0	2min 30s
10^{-6}	76	4147*404.89779*	15.6	2min 41s
10^{-8}	78	4147385.53*899*	16.9	2min 54s
10^{-10}	78	4147385.53906	18.1	3min 02s
		`filter2D`		
10^{-2}	194	32566.*5899785*	7.0	4.3s
10^{-4}	195	32566.*5896021*	10.6	5.0s
10^{-6}	199	32566.8479*516*	13.7	5.7s
10^{-8}	199	32566.8479722	16.9	6.8s
10^{-10}	199	32566.8479722	20.1	7.1s
		`rajat19`		
10^{-2}	484	26985.*7363375*	7.3	2.8s
10^{-4}	487	26985.*7841332*	12.4	3.5s
10^{-6}	494	26985.9640*290*	13.7	3.7s
10^{-8}	494	26985.9640071	14.6	3.7s
10^{-10}	494	26985.9640071	15.2	3.8s

maximum accuracy of the quantity

$$\left\| \frac{\nabla f(\boldsymbol{x}_{p+1})}{f(\boldsymbol{x}_{p+1})} - \frac{\nabla f(\boldsymbol{x}_p)}{f(\boldsymbol{x}_p)} \right\|_\infty \tag{14.6}$$

(where $\|\cdot\|_\infty$ is the maximum norm) is smaller than the unit roundoff $\mathbf{u} = 2^{-53} \approx 1.11 \times 10^{-16}$. In other words, $\boldsymbol{x}_p$ is eventually stationary to machine precision.

The separation of the gradient curves observed for `1138_bus` and `tuma2` indicates that different contractions may yield different trajectories in the objective landscape. However, when γ is sufficiently small (below 10^{-6} for `1138 bus` and 10^{-8} for `tuma2`), the gradient curve does not change and hence represents the behavior of exact QMM until maximum accuracy is attained. Looking at the short run, we see that PCG-QMM behaves similarly to exact QMM until the outer termination criterion is satisfied.

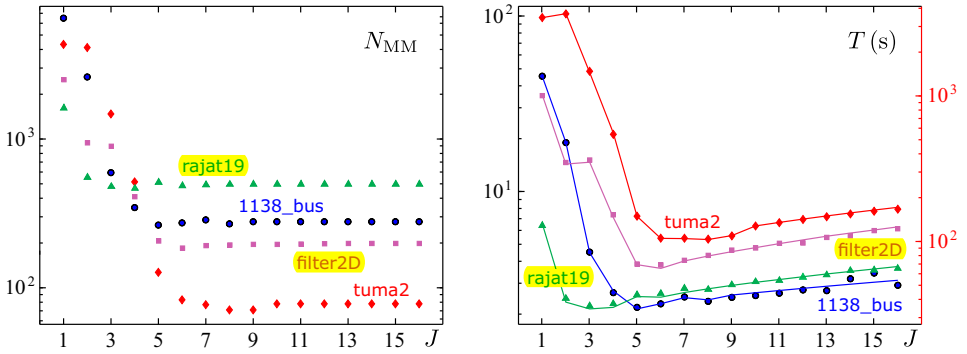

FIGURE 5. Behavior of the PCG-QMM graph layout algorithm with a fixed number J of PCG iterations per outer iteration: number of outer iterations (left) and running time (right) versus J for an outer termination tolerance of 10^{-6}.

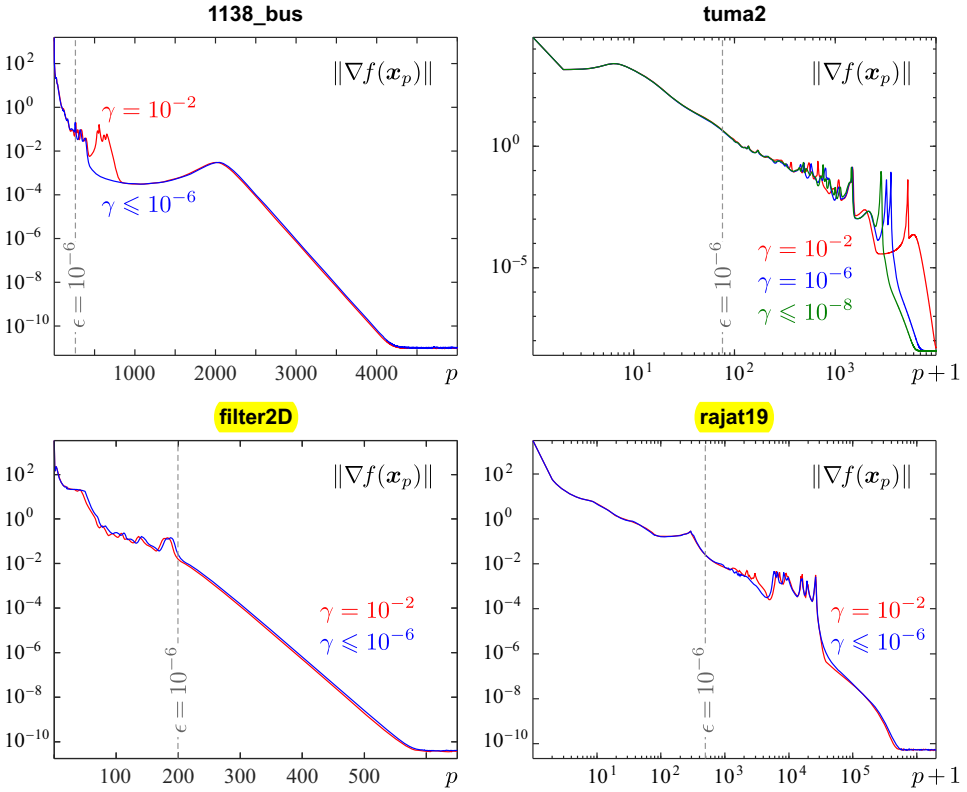

FIGURE 6. Graph layout: norm of the gradient versus number of outer iterations (stopping delay $k = 5$).

We now distinguish three types of solutions:

(i) the *practical solutions* $\widehat{\boldsymbol{x}} := \boldsymbol{x}_{N_{\mathrm{MM}}}$ obtained using both the inner and outer termination criteria;

(ii) the *limit solutions*, denoted by $\widehat{\boldsymbol{x}}^*$, obtained by continuing the outer iterations up to maximum accuracy;

(iii) the *double-limit solutions*, denoted by $\widehat{\boldsymbol{x}}^{**}$ and independent of γ, obtained by letting the PCG solver reach its accuracy limit at every outer iteration and by continuing the outer iterations up to maximum accuracy.

To assess the visual difference between two solutions, we use the *Tucker distance*

$$\tau(\boldsymbol{x},\boldsymbol{y}) := 1 - \frac{\sum_{i,j} \varrho_{ij}(\boldsymbol{x})\varrho_{ij}(\boldsymbol{y})}{\big(\sum_{i,j} \varrho_{ij}(\boldsymbol{x})^2\big)^{1/2}\big(\sum_{i,j} \varrho_{ij}(\boldsymbol{y})^2\big)^{1/2}} \in [0,1], \tag{14.7a}$$

$$\varrho_{ij}(\boldsymbol{x}) := \big\|\boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)}\big\|/d_{ij}. \tag{14.7b}$$

(The Euclidean distance $\|\boldsymbol{x}-\boldsymbol{y}\|$ is inadequate because the objective is invariant to proper and improper rotations of the layout around the origin.)

Figures 7 and 8 plot the normalized objective

$$f^*(\boldsymbol{x}_p) := \frac{f(\boldsymbol{x}_p) - f(\widehat{\boldsymbol{x}}^*)}{f(\boldsymbol{x}_0) - f(\widehat{\boldsymbol{x}}^*)} \tag{14.8}$$

and the distance to the limit solution, $\|\boldsymbol{x}_p - \widehat{\boldsymbol{x}}^*\|$. The objective decreases monotonically and plateaus between gradient peaks. In other words, the trajectory of the iterates switches between nearly flat regions in the vicinity of $\widehat{\boldsymbol{x}}^*$, which can be interpreted as aggregations of saddle or small-curvature points around minimizers. The distance to $\widehat{\boldsymbol{x}}^*$ eventually decays exponentially, indicating R-linear convergence.

Table 3 compares the limit, double-limit, and practical solutions: the second and third columns give the Euclidean and Tucker distances between $\boldsymbol{x}^{\widehat{}\,*}$ and $\boldsymbol{x}^{\widehat{}\,**}$, and the last two columns compare $\boldsymbol{x}^{\widehat{}}$ and $\boldsymbol{x}^{\widehat{}\,**}$ in terms of relative objective difference and Tucker distance. Although the values of $\|\boldsymbol{x}^{\widehat{}\,*} - \boldsymbol{x}^{\widehat{}\,**}\|$ for $\gamma = 10^{-2}$ and $\gamma = 10^{-6}$ differ significantly, the tucker distance $\tau(\boldsymbol{x}^{\widehat{}\,*}, \boldsymbol{x}^{\widehat{}\,**})$ is smaller than 2×10^{-3}, meaning that the limit solutions for $\gamma \leqslant 10^{-2}$ are visually indistinguishable from $\widehat{\boldsymbol{x}}^{**}$ (up to rotation). Furthermore, the relative objective difference and the Tucker distance between $\boldsymbol{x}^{\widehat{}}$ and $\boldsymbol{x}^{\widehat{}\,**}$ are smaller than 3×10^{-4} and 5×10^{-3}, respectively. So decreasing γ below 10^{-2} or decreasing the outer termination tolerance below 10^{-6} does not improve the layout aesthetics.

14.1.3. *Robustness to the parameters of the PCG solver.* We set the outer termination tolerance ϵ to 10^{-6} and look at the behavior of the PCG-QMM algorithm for a stopping delay ranging from 1 to 32 and a contraction number between 10^{-8} and 10^{-1}. Table 4 gives the ranges of (i) the relative objective difference between the pratical and double-limit solutions $\boldsymbol{x}^{\widehat{}}$ and $\boldsymbol{x}^{\widehat{}\,**}$, (ii) the Tucker distance between $\widehat{\boldsymbol{x}}$ and $\boldsymbol{x}^{\widehat{}\,**}$, (iii) the running time, and (iv) the number of outer iterations. We see that the practical solutions are very close to the double limit solutions, and therefore not sensitive to the parameters of the PCG solver. Furthermore, the maximum-to-minimum ratios of the running time and of the number of outer it-erations (less than 4 and 2, respectively) are small relative to the ranges of k and γ. The running time is maximal for $(k, \gamma) = (32, 10^{-8})$ and minimal or close to minimal for $(k, \gamma) = (2, 10^{-2})$, and the number of outer iterations is maximal for $(k, \gamma) = (1, 10^{-1})$ and stabilizes as k increases and/or γ decreases.

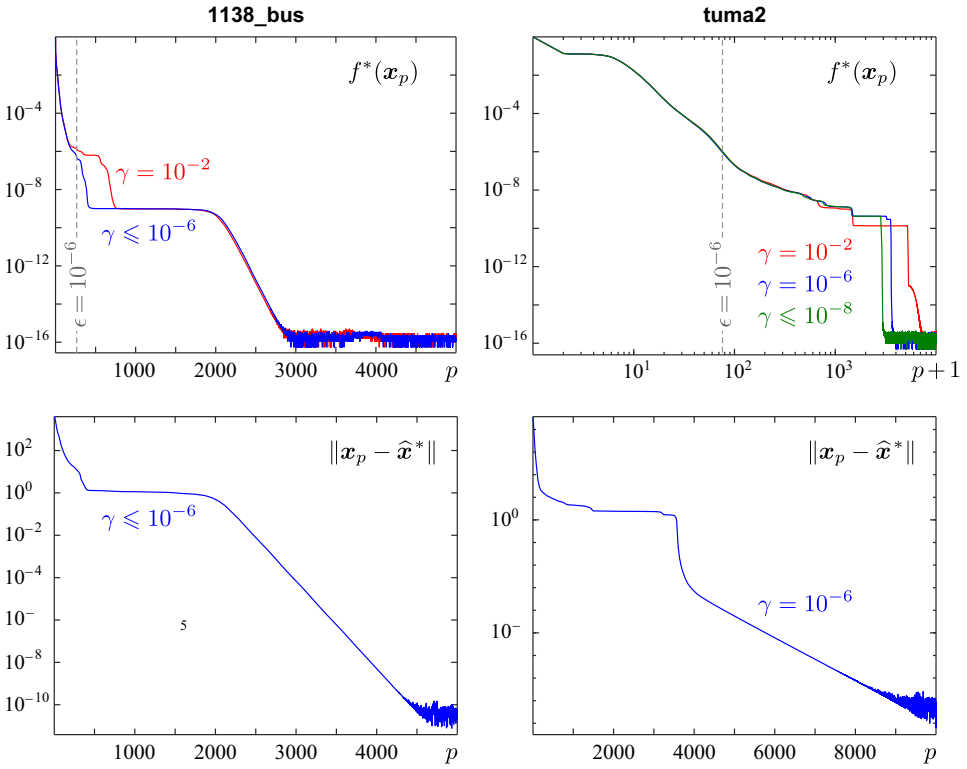

FIGURE 7. Graph layout: normalized objective (top) and distance to the limit (bottom) versus number of outer iterations (stopping delay $k = 5$).

14.1.4. *Reliability of the PCG stopping criterion.* Based on Lemma 12.1, we define the contraction from $\boldsymbol{x}$ to $\boldsymbol{y}$ by

$$\Gamma(\boldsymbol{y}\,|\boldsymbol{x}) := \left(\frac{E(\boldsymbol{y}\,|\boldsymbol{x})}{E(\boldsymbol{x}|\boldsymbol{x})}\right)^2. \tag{14.9}$$

Then $\boldsymbol{y} \in \boldsymbol{\Phi}_\gamma(\boldsymbol{x})$ if and only if $\Gamma(\boldsymbol{y}\,|\boldsymbol{x}) \leqslant \gamma$, and the PCG stopping test (12.24) is designed to guarantee that

$$\Gamma(\boldsymbol{y}_{j_p+1-k}\,|\,\boldsymbol{x}_p) \lesssim \gamma, \tag{14.10}$$

with the quality of the approximation increasing with k. (Recall that $\boldsymbol{x}_{p+1} = \boldsymbol{y}_{j_p+1}$, so the outer iterates are k PCG iterations beyond those actually needed to obtain a contraction close to γ.) From (12.23), this test is numerically stable if

$$n^2\mathbf{u}\max\left(\kappa/\sqrt{n},\, \kappa^\dagger\right) \ll 1, \tag{14.11}$$

where $n = 2(m-1)$ and κ and $\kappa^\dagger$ are respectively the condition numbers of the inner system matrix and the preconditioner. The quantity on the left-hand side of (14.11) is of order 10^{-5} for `tuma2` and 10^{-8} for `1138 bus`, `filter2D` and `rajat19` ($\kappa/\sqrt{n} > \kappa^\dagger$ in all cases). Therefore, failing to satisfy (14.10) is not due to finite precision arithmetic but to the fact that k is too small.

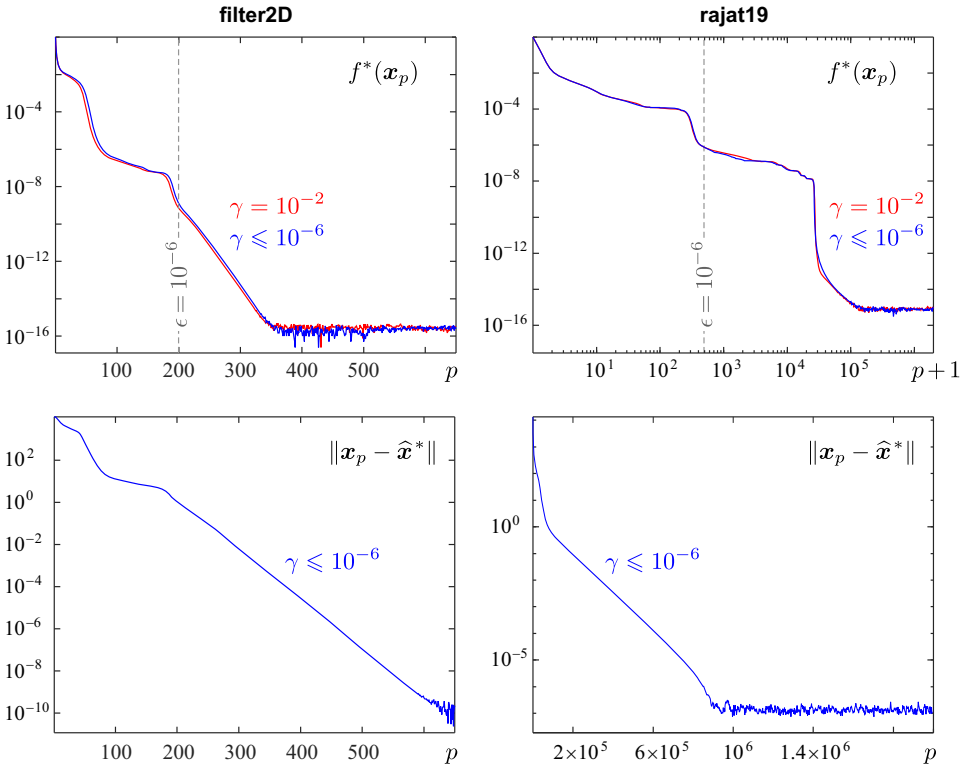

FIGURE 8. Graph layout: normalized objective (top) and distance to the limit (bottom) versus number of outer iterations (stopping delay $k = 5$).

We distinguish between the *inner contraction* $\Gamma(\boldsymbol{y}_{j_p+1-k}\,|\,\boldsymbol{x}_p)$ and the *outer contraction* $\Gamma(\boldsymbol{x}_{p+1}\,|\,\boldsymbol{x}_p)$. We denote by Γ_{inner} and Γ_{outer} their respective maximum values over the course of the algorithm (before reaching maximum accuracy). As k increases, we expect Γ_{inner} to get closer to γ and Γ_{outer} to decrease. Table 5 gives the maximum inner and outer contractions for different values of k when $\gamma = 10^{-6}$. We see that the inner contraction is close to γ for k large enough ($k \geqslant 6$ for `1138 bus`, $k \geqslant 7$ for `tuma2` and `filter2D`, and $k \geqslant 5$ for `rajat19`) and that Γ_{outer} is rapidly decreasing with increasing k.

14.2. **X-ray tomography.** The problem considered here is to reconstruct the middle cross-section of a walnut shown in Figure 9(a) from the limited fan-beam data in Figure 9(b). These data are publicly available [64]; they consist of 120 projections evenly spaced over 360°, each containing 328 measurements with a step size of 0.35 mm. The *ground truth* $\boldsymbol{x}^\star$ in Figure 9(a) is a 328×328 downsampled version of a 2296×2296 filtered back-projection reconstruction from a 1200×2296 sinogram (0.3° angular step, $50\,\mu$m detector size). The observation matrix $\boldsymbol{D}$ has $m = 39360$ rows and $n = 328^2$ columns. Its density (the percentage ratio of its number of nonzero entries to its total number of entries) is about 0.37%.

TABLE 3. Graph layout: comparison of the limit, double-limit, and practical solutions ($\boldsymbol{x}^{\widehat{}*}$, $\boldsymbol{x}^{\widehat{}**}$, and $\boldsymbol{x}^{\widehat{}}$, respectively). The stopping delay (for $\boldsymbol{x}^{\widehat{}*}$ and $\boldsymbol{x}^{\widehat{}}$) is $k = 5$ and the outer termination tolerance (for $\boldsymbol{x}^{\widehat{}}$) is $\epsilon = 10^{-6}$.

γ	$\|\boldsymbol{x}^{\widehat{}*} - \boldsymbol{x}^{\widehat{}**}\|$	$\tau(\boldsymbol{x}^{\widehat{}*}, \boldsymbol{x}^{\widehat{}**})$	$f(\boldsymbol{x}^{\widehat{}})/f(\boldsymbol{x}^{\widehat{}**}) - 1$	$\tau(\boldsymbol{x}^{\widehat{}}, \boldsymbol{x}^{\widehat{}**})$
		`1138_bus`		
10^{-2}	21.1	4.6×10^{-5}	3.3×10^{-5}	1.3×10^{-4}
10^{-6}	3.8×10^{-4}	2.2×10^{-16}	1.9×10^{-5}	3.0×10^{-5}
		`tuma2`		
10^{-2}	156.5	6.8×10^{-5}	5.8×10^{-5}	8.6×10^{-4}
10^{-6}	15.3	2.1×10^{-5}	6.2×10^{-5}	8.5×10^{-4}
		`filter2D`		
10^{-2}	375.6	1.4×10^{-5}	7.6×10^{-7}	1.5×10^{-5}
10^{-6}	0.76	0	3.5×10^{-7}	9.9×10^{-7}
		`rajat19`		
10^{-2}	469.3	1.6×10^{-3}	2.6×10^{-4}	4.3×10^{-3}
10^{-6}	4.2×10^{-3}	0	2.7×10^{-4}	4.0×10^{-3}

TABLE 4. Overall behavior of the PCG-QMM graph layout algorithm for $(k, \gamma) \in \{1, \ldots, 32\} \times [10^{-8}, 10^{-1}]$ (outer termination tolerance $\epsilon = 10^{-6}$).

	$f(\boldsymbol{x}^{\widehat{}})/f(\boldsymbol{x}^{\widehat{}**}) - 1$	$\tau(\boldsymbol{x}^{\widehat{}}, \boldsymbol{x}^{\widehat{}**})$	T (s)	N_{MM}
`1138_bus`	$[0.0031, 1.0] \times 10^{-4}$	$[0.30, 4.3] \times 10^{-4}$	$[2.1, 5.6]$	$[264, 461]$
`tuma2`	$[0.67, 6.3] \times 10^{-5}$	$[4.0, 9.3] \times 10^{-4}$	$[93, 371]$	$[73, 139]$
`filter2D`	$[0.019, 1.2] \times 10^{-5}$	$[0.048, 1.5] \times 10^{-5}$	$[3.9, 13]$	$[193, 227]$
`rajat19`	$[1.6, 3.0] \times 10^{-4}$	$[4.0, 6.1] \times 10^{-3}$	$[2.5, 7.5]$	$[479, 574]$

We look for solutions that minimize

$$f(\boldsymbol{x}) := \|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{d}\|^2 + \lambda \sum_{j=1}^{l} \theta\big(\|\boldsymbol{R}_j \boldsymbol{x}\|/\delta\big), \tag{14.12}$$

where $\lambda > 0$ controls the regularization strength and $\delta > 0$ adjusts the scale of the operator $\{\boldsymbol{R}_j\}_j$. This objective has the form (13.22) with $M = m + l$ and

$$(\boldsymbol{A}_i, \boldsymbol{a}_i, \theta_i(t)) = \begin{cases} (\boldsymbol{D}(i,:), [\boldsymbol{d}]_i, t^2) \text{ if } i \leqslant m, (\boldsymbol{R}_i \\ -m, \boldsymbol{0}, \lambda\theta(t/\delta)) \text{ otherwise.} \end{cases} \tag{14.13}$$

The experiments below illustrate the behavior of the PCGLS-QMM algorithm in convex and nonconvex settings. In the convex case,

$$\theta(u) = (1 + u^2)^{1/2} - 1 =: \theta_{\mathrm{MS}}(u) \tag{14.14}$$

TABLE 5. Graph layout: maximum inner and outer contractions versus stopping delay (target contraction $\gamma = 10^{-6}$).

	`1138_bus`		`tuma2`	
k	Γ_{inner}	Γ_{outer}	Γ_{inner}	Γ_{outer}
3	1.15×10^{-5}	1.05×10^{-5}	1.76×10^{-4}	1.75×10^{-4}
4	2.85×10^{-6}	1.87×10^{-6}	4.61×10^{-5}	4.51×10^{-5}
5	1.22×10^{-6}	2.76×10^{-7}	8.83×10^{-6}	7.85×10^{-6}
6	1.00×10^{-6}	2.07×10^{-8}	2.04×10^{-6}	1.10×10^{-6}
7	9.94×10^{-7}	6.44×10^{-10}	1.01×10^{-6}	4.46×10^{-8}
8	9.94×10^{-7}	2.73×10^{-11}	9.99×10^{-7}	3.59×10^{-9}
	`filter2D`		`rajat19`	
k	Γ_{inner}	Γ_{outer}	Γ_{inner}	Γ_{outer}
3	2.32×10^{-5}	2.23×10^{-5}	3.79×10^{-6}	2.79×10^{-6}
4	9.91×10^{-6}	8.93×10^{-6}	1.41×10^{-6}	4.06×10^{-7}
5	4.18×10^{-6}	3.25×10^{-6}	1.05×10^{-6}	4.67×10^{-8}
6	1.96×10^{-6}	9.95×10^{-7}	1.02×10^{-6}	5.76×10^{-10}
7	1.07×10^{-6}	1.66×10^{-7}	1.01×10^{-6}	8.09×10^{-12}
8	1.01×10^{-6}	2.32×10^{-8}	1.01×10^{-6}	2.10×10^{-14}

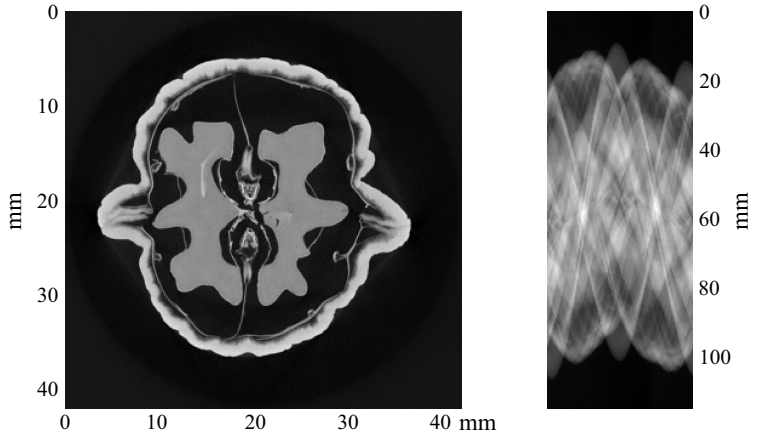

FIGURE 9. Tomography problem: cross-section of a walnut (left) and sinogram data (right).

(the function of minimal surfaces [65]), and $\{\boldsymbol{R}_j\}_j$ is the usual spatial-gradient operator [53, Section 5.2]. The intersection of the null spaces of the $\boldsymbol{R}_j$'s is the set of constant images, and since the only constant image in null($\boldsymbol{D}$) is $\mathbf{0}$, Theorem 13.3 guarantees R-linear convergence to the global minimizer of f. In the nonconvex case,

$$\theta(u) = \ln(1 + u^2) =: \theta_{\text{LE}}(u) \tag{14.15}$$

(the Lorentzian error function [66]), and $\{\boldsymbol{R}_j\}_j$ consists of the high-pass operators of a two-level framelet system generated from piecewise linear filters [67]. Since θ_{LE} is analytic, Theorem 13.2 predicts R-linear or -sublinear convergence to a stationary point of f.

14.2.1. *Quality measures and parameter setting.* The quality of the computed solutions is assessed with the peak signal-to-noise ratio and with structural dissimilarity. The peak signal-to-noise ratio (in dB) of a reconstruction $\boldsymbol{x}$ is defined as

$$\mathrm{PSNR} := 20\log_{10}\biggl(\frac{\sqrt{n}\,\boldsymbol{x}^\star_{\mathrm{pp}}}{\|\boldsymbol{x}-\boldsymbol{x}^\star\|}\biggr),$$

where $\boldsymbol{x}^\star_{\mathrm{pp}}$ is the peak-to-peak amplitude of the ground truth. The structural dissimilarity is $\mathrm{SD} := 1 - \mathrm{SSIM}$, where SSIM is the popular structural similarity index [68].

The free parameters λ and δ are selected as follows. Given the percentage area α occupied by air in the ground truth, the αth percentile P_α of $\{\|\boldsymbol{R}_j\boldsymbol{x}^\star\|\}_j$ can be interpreted as the minimum magnitude for a regularization vector (or scalar) $\boldsymbol{R}_j\boldsymbol{x}^\star$ to be considered significant; it is about 1.2×10^{-3} for the gradient operator and about 2.1×10^{-4} for the framelet operator. We set $\delta = P_\alpha/10$ in the former case (so the convex regularizer is a smooth approximation to the ℓ_1-norm of the gradient) and $\delta = P_\alpha$ in the latter (so the significant framelet coefficients are preserved). We then adjust λ to achieve near optimal PSNR (about 26.4 dB in both cases).

We use the following continuation sequences. In the convex case, the relaxed objectives f_p are the same as f but with $\theta(\cdot/\delta)$ replaced by $(\delta_p/\delta)\theta_{\mathrm{MS}}(\cdot/\delta_p)$, where $(\delta_p)_{0\leqslant p\leqslant q}$ decreases linearly from 100δ to δ. In the nonconvex case, f_p is obtained by replacing θ by $\mu_p\theta_{\mathrm{LE}} + (1-\mu_p)\theta_{\mathrm{MS}}$, where $(\mu_p)_{0\leqslant p\leqslant q}$ increases linearly from 0 to 1. The length q of the continuation sequences is set to 50.

The starting point is always the zero image and we use the Jacobi preconditioner

$$\boldsymbol{M}(\boldsymbol{x}) = \mathrm{diag}(\boldsymbol{W}(\boldsymbol{x})) = \mathrm{diag}\bigl(\|\boldsymbol{A}(:,1)\|^2_{\boldsymbol{Q}(\boldsymbol{x})},\dots,\|\boldsymbol{A}(:,n)\|^2_{\boldsymbol{Q}(\boldsymbol{x})}\bigr). \tag{14.16}$$

As in the previous experiments, we distinguish between practical, limit, and double-limit solutions (denoted by $\widehat{\boldsymbol{x}}$, $\widehat{\boldsymbol{x}}^*$, and $\widehat{\boldsymbol{x}}^{**}$, respectively).

14.2.2. *Robustness to the parameters of the PCG solver.* We set the outer termination tolerance ϵ to 10^{-5}. Figure 10 shows the reconstructions obtained for $(k,\gamma) = (5, 10^{-4})$. The nonconvex regularizer yields a sharper image, as expected by design. Table 6 gives the mean and maximum relative difference of the objective value, PSNR and structural dissimilarity of the practical solutions for a stopping delay ranging from 1 to 32 and a contraction number between 10^{-8} and 10^{-1}. Looking at the relative differences, we see that the reconstructions are not sensitive to the parameters of the PCG solver. In the nonconvex case, this robustness suggests that the algorithm always ends up in the bottom of the same basin.

The running time is approximately affine in both k and $-\log_{10}(\gamma)$ (the order of magnitude of the contraction number) as well as in the density of the factor $\boldsymbol{A}(\boldsymbol{x})$ of the weighting matrix. In the convex case, the running time ranges from about 2 minutes for $(k,\,\gamma) = (1,\,0.1)$ to about 10 minutes for $(k,\,\gamma) = (32,\,10^{-8})$, and the number N_{MM} of outer iterations is between 220 and 225. In the nonconvex case, the running time ranges from 15 to 52 minutes, and N_{MM} is between 360 and 400 when $k \geqslant 4$ and goes up to 660 for $(k,\gamma) = (1, 0.1)$.

TABLE 6. Overall behavior of the PCGLS-QMM reconstruction algorithm for $(k, \gamma) \in \{1, \ldots, 32\} \times [10^{-8}, 10^{-1}]$ (outer termination tolerance $\epsilon = 10^{-5}$).

	Convex regularizer		Nonconvex regularizer	
	Mean	Max. rel. diff.	Mean	Max. rel. diff.
$f(\widehat{\boldsymbol{x}})$	15.20	1.2×10^{-10}	14.19	9.6×10^{-6}
PSNR (dB)	26.44	1.3×10^{-7}	26.36	4.6×10^{-5}
SD	2.40×10^{-7}	3.5×10^{-7}	2.36×10^{-7}	1.5×10^{-4}

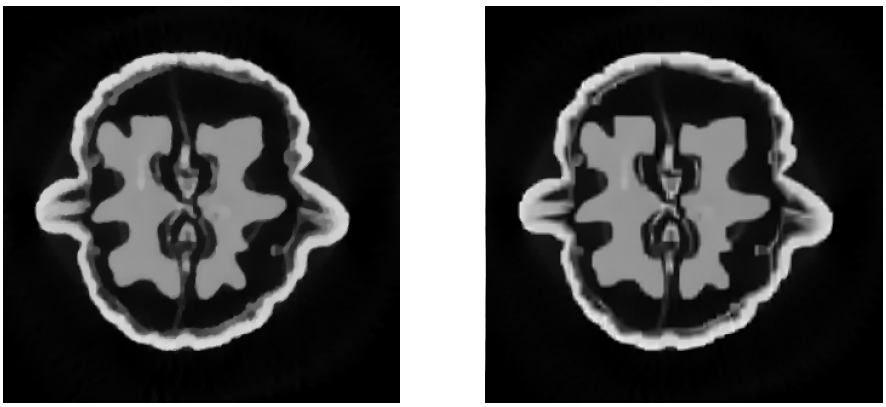

FIGURE 10. Reconstructions obtained using the convex gradient regularizer (left) and the nonconvex framelet regularizer (right).

14.2.3. *Behavior in the long run.* Figure 11 plots the norm of the gradient, the normalized objective (see (14.8)), and the Euclidean distance to the limit for different values of the stopping delay and the contraction number (the vertical dashed lines indicate the number of outer iterations to termination when $\epsilon = 10^{-5}$). We make the same observations as for graph layout: the iterates at maximum accuracy are stationary to machine precision, the objective decreases monotonically, and the convergence is R-linear. In the convex case, PCGLS-QMM behaves similarly to exact QMM independently of k when $\gamma \leqslant 0.1$, while in the nonconvex case the trajectory of the iterates is stable when $k \geqslant 5$ and $\gamma \leqslant 10^{-4}$. We also see that decreasing γ does not necessarily improve the convergence rate.

Table 7 compares the practical solutions obtained for $(k, \gamma) = (1, 0.1)$ with the double-limit solutions. Their differences in terms of Euclidean distance, objective, PSNR, and structural dissimilarity are negligible (and they are indeed visually indistinguishable from the reconstructions shown in Figure 10). So again we conclude that there is no point in choosing γ or ϵ too small, be it in terms of execution time or solution quality.

14.2.4. *The PCG stopping test at work.* Figure 12 plots the inner and outer contractions $\Gamma(\boldsymbol{y}_{j_p+1-k} \,|\, \boldsymbol{x}_p)$ and $\Gamma(\boldsymbol{x}_{p+1} \,|\, \boldsymbol{x}_p)$ (as defined in Section 14.1.4) for different values of the stopping delay and a target $\gamma = 10^{-4}$. As expected, the maximum inner contraction is close to γ for k large enough ($k \geqslant 20$ in the convex case and $k \geqslant 5$

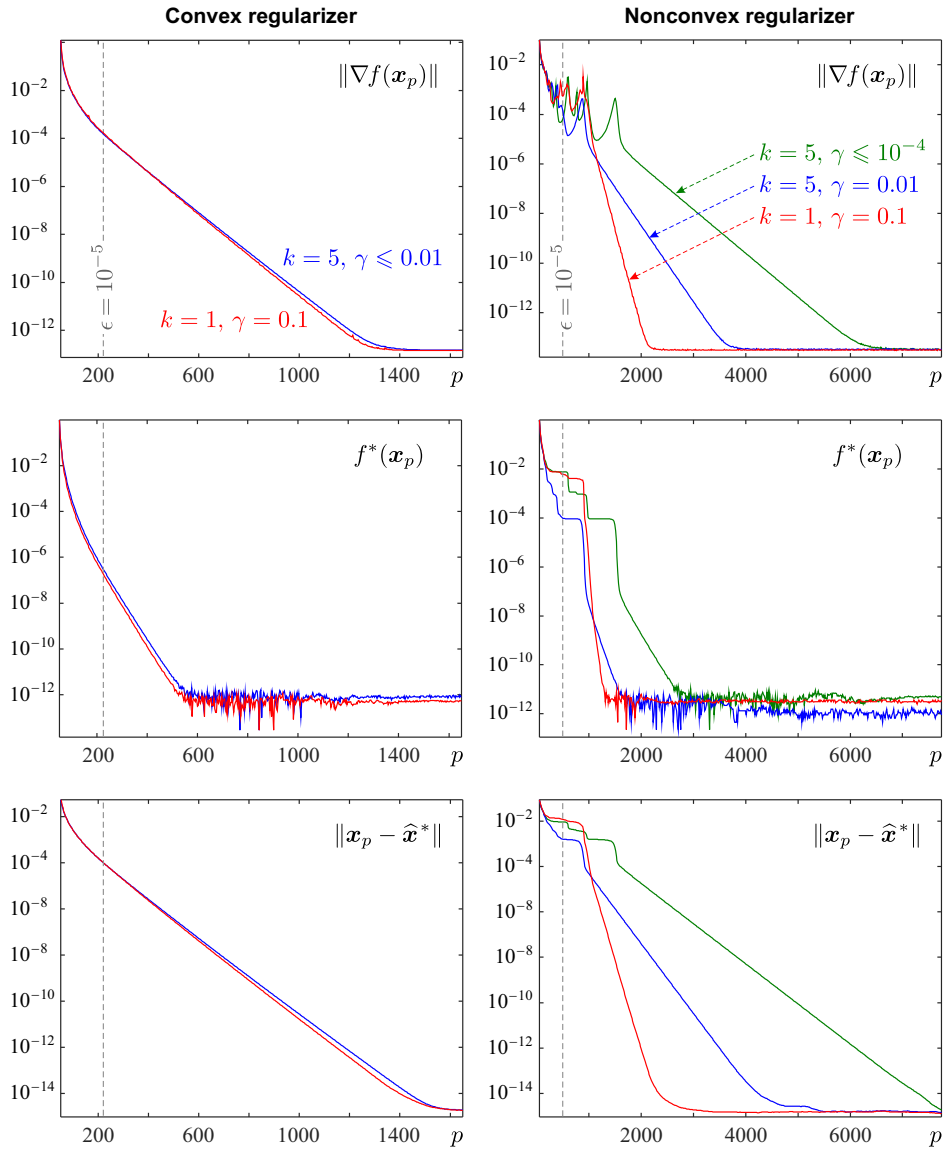

FIGURE 11. Long-run behavior of the PCGLS-QMM reconstruction algorithm. From top to bottom: norm of the gradient, normalized objective, and distance to the limit versus number of outer iterations.

in the nonconvex case), and the ratio of the outer to inner contraction decreases rapidly with increasing k.

We notice that the minimum suitable value of k for matching γ is larger in the convex case. This indicates that, along the iterate trajectories, the spectra of the surrogates (that is, the distributions of the eigenvalues of the inner system matrices) are more spread out for the convex objective than for the nonconvex one. Indeed, we have seen in Section 12.2 that the larger k, the better the approximation

TABLE 7. Tomographic reconstruction: comparison of the practical solutions obtained for $(k, \gamma) = (1, 0.1)$ with the double-limit solutions.

	Convex regularizer	Nonconvex regularizer
$\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}^{**}\| / \|\widehat{\boldsymbol{x}}^{**}\|$	1.1×10^{-5}	2.7×10^{-3}
$\|f(\widehat{\boldsymbol{x}})/f(\widehat{\boldsymbol{x}}^{**}) - 1\|$	7.9×10^{-10}	5.3×10^{-6}
$\|\mathrm{PSNR}(\widehat{\boldsymbol{x}})/\mathrm{PSNR}(\widehat{\boldsymbol{x}}^{**}) - 1\|$	6.4×10^{-7}	2.5×10^{-6}
$\|\mathrm{SD}(\widehat{\boldsymbol{x}})/\mathrm{SD}(\widehat{\boldsymbol{x}}^{**}) - 1\|$	2.0×10^{-6}	1.7×10^{-4}

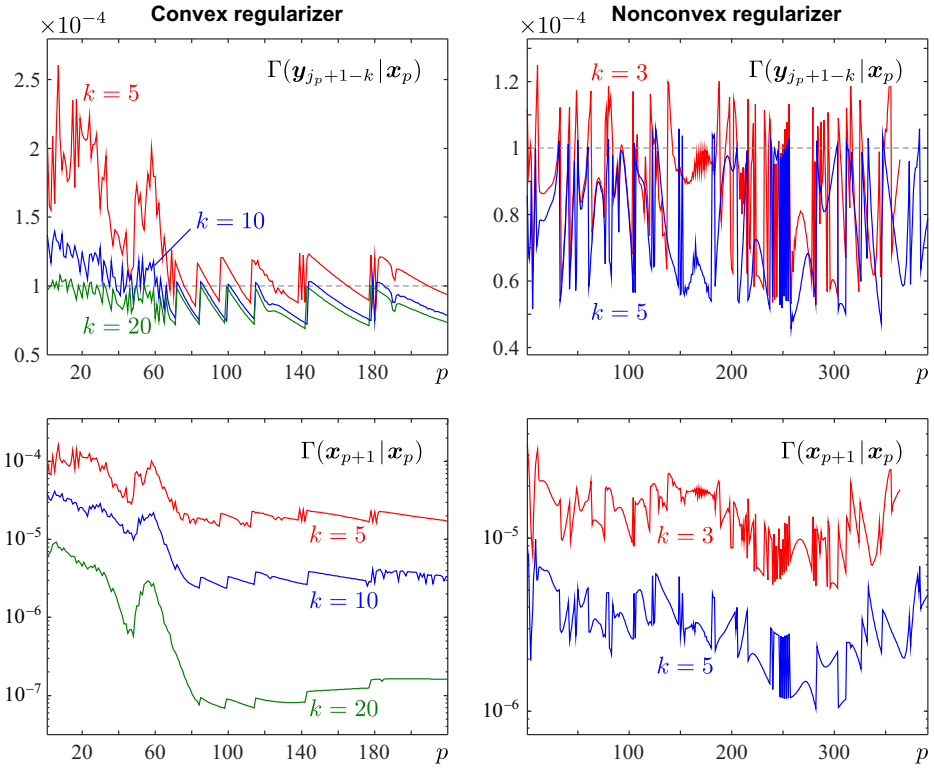

FIGURE 12. Tomographic reconstruction: inner and outer contractions (top and bottom, respectively) versus number of outer iterations. The target contraction is $\gamma = 10^{-4}$.

(12.10) of the inner error energy norm, and hence the closer the inner contraction to γ. At the same time, the error bound (12.13) not only shows that the quality of this approximation depends on the spectrum of the inner system matrix, but also gives insight into what qualifies as a favorable spectrum, namely tightly clustered eigenvalues away from the origin, and what constitutes an unfavorable one, namely widely spread out eigenvalues (see [35, Section 3.1]). Therefore, the more scattered the eigenvalues, the larger k must be for matching γ.

TABLE 8. Comparison of PCGLS-QMM(k, γ) with nonlinear CG-related methods: running time T and number of iterations N for an outer termination tolerance of 10^{-5}.

	Convex regularizer		Nonconvex regularizer	
	T	N	T	N
TN	19min 51s	381	3h 03min 31s	502
CG-PR+	20min 02s	400	3h 26min 56s	557
L-BFGS	2min 28s	261	27min 51s	355
PCGLS-QMM(1, 10^{-1})	2min 05s	225	14min 55s	324
PCGLS-QMM(5, 10^{-2})	2min 30s	220	22min 16s	425

TABLE 9. Comparison of PCGLS-QMM(k, γ) with nonlinear CG-related methods in the long run: minimum gradient norm attained and number of iterations to stagnation N^*.

	Convex regularizer		Nonconvex regularizer	
	$\min_p \|\nabla f(\boldsymbol{x}_p)\|$	N^*	$\min_p \|\nabla f(\boldsymbol{x}_p)\|$	N^*
TN	2.9×10^{-6}	483	1.3×10^{-6}	582
CG-PR+	3.1×10^{-6}	522	6.2×10^{-6}	662
L-BFGS	2.6×10^{-6}	399	1.1×10^{-6}	589
PCGLS-QMM(1, 10^{-1})	1.5×10^{-13}	1400	3.3×10^{-14}	2300
PCGLS-QMM(5, 10^{-2})	1.4×10^{-13}	1500	3.1×10^{-14}	4000

14.2.5. *Comparison with other nonlinear CG-related algorithms.* Finally, we compare PCGLS-QMM with three nonlinear CG-related methods: truncated Newton (TN) [69], Polak-Ribi`ere conjugate gradient (CG-PR+) [70], and L-BFGS [71]. These three algorithms are major tools for large-scale optimization, but they do not come with as strong convergence guarantees as PCGLS-QMM. The limited memory parameter of L-BFGS is set to the recommended value of 5 [72], and we consider PCGLS-QMM in fast and normal settings (namely, (k, γ) = (1, 0.1) and (5, 10^{-2}), respectively). For an outer termination tolerance of 10^{-5}, TN, CG-PR+ and L-BFGS all produce similar reconstructions to PCGLS-QMM: the objective, PSNR and SD are the same to four significant digits in both the convex and non-convex cases. However, as Table 8 shows, there are significant differences in running time: PCGLS-QMM is 7 to 14 times faster than TN and CG-PR+, and up to twice as fast as L-BFGS. Furthermore, in the long run, the accuracy of TN, CG-PR+ and L-BFGS plateaus prematurely compared to PCGLS-QMM. This is illustrated in Table 9, which gives the number of iterations to and the value of the minimum gradient norm: the gradient norm of the iterates of TN, CG-PR+ and L-BFGS plateaus above 10^{-6} after 400–700 iterations, while PCGLS-QMM plateaus below 1.5×10^{-13} after 1400–4000 iterations.

References

[1] D. Hunter and K. Lange, *A tutorial on MM algorithms*, Amer. Statist. **58** (2004), no. 1, 30–37.

[2] T. Wu and K. Lange, *The MM alternative to EM*, Statist. Sci. **25** (2010), no. 4, 492–505.

[3] K. Lange, *MM optimization algorithms*, SIAM, 2016.

[4] Y. Sun, P. Babu, and D. Palomar, *Majorization-minimization algorithms in signal processing, communications, and machine learning*, IEEE Trans. Signal Processing **65** (2017), no. 3, 794–816.

[5] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, *Sparse multinomial logistic regression: fast algorithms and generalization bounds*, IEEE Trans. Pattern Anal. Machine Intell. **27** (2005), no. 6, 957–968.

[6] D. Ba, B. Babadi, P. Purdon, and E. Brown, *Convergence and stability of iteratively re-weighted least squares algorithms*, IEEE Trans. Signal Processing **62** (2014), no. 1, 183–195.

[7] T. Qiu, P. Babu, and D. Palomar, *PRIME: phase retrieval via majorization-minimization*, IEEE Trans. Signal Processing **64** (2016), no. 19, 5174–5186.

[8] P. Oğuz-Ekim, J. Gomes, J. Xavier, and P. Oliveira, *Robust localization of nodes and time-recursive tracking in sensor networks using noisy range measurements* **59** (2011), no. 8, 3930–3942.

[9] I. Borg and P. Groenen, *Modern multidimensional scaling*, Springer, 2009.

[10] M. Robini, F. Yang, and Y. Zhu, *Inexact half-quadratic optimization for linear inverse problems*, SIAM J. Imaging Sci. **11** (2018), no. 2, 1078–1133.

[11] H. Attouch, J. Bolte, and B. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program. Ser. A **137** (2013), no. 1, 91–129.

[12] M. Razaviyayn, M. Hong, and Z.-Q. Luo, *A unified convergence analysis of block successive minimization methods for nonsmooth optimization*, SIAM J. Optim. **23** (2013), no. 2, 1126–1153.

[13] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, *Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function*, J. Optim. Theory Appl. **162** (2014), no. 1, 107–132.

[14] J. Mairal, *Incremental majorization-minimization optimization with application to large-scale machine learning*, SIAM J. Optim. **25** (2015), no. 2, 829–855.

[15] J. Bolte and E. Pauwels, *Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs*, Math. Oper. Res. **41** (2016), no. 2, 442–465.

[16] A. Beck and D. Pan, *Convergence of an inexact majorization-minimization method for solving a class of composite optimization problems*, Large-scale and distributed optimization, Lecture Notes in Math. **2227** (2018), 375–410.

[17] A. Ostrowski, *Solution of equations in Euclidean and Banach spaces*, Academic Press, 1973.

[18] P. Ciarlet, *Introduction to numerical linear algebra and optimisation*, Cambridge University Press, 1989.

[19] S. Łojasiewicz, *Ensembles semi-analytiques*, Institut des Hautes Études Scientifiques, 1965.

[20] E. Bierstone and P. Milman, *Semianalytic and subanalytic sets*, Publ. Math. Inst. Hautes Études Sci. **67** (1988), 5–42.

[21] ———, *Canonical desingularization in characteristic zero by blowing up the maximum strata of a local invariant* **128** (1997), no. 2, 207–302.

[22] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Ann. Inst. Fourier **48** (1998), no. 3, 769–783.

[23] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Math. J. **84** (1996), no. 2, 497–540.

[24] L. van den Dries, *Tame topology and o-minimal structures*, Cambridge University Press, 1998.

[25] ———, *O-minimal structures and real analytic geometry*, Current developments in mathematics, International Press, 1998, pp. 105–152.

[26] J.-P. Rolin, P. Speissegger, and A. Wilkie, *Quasianalytic Denjoy-Carleman classes and o-minimality*, J. Amer. Math. Soc. **16** (2003), no. 4, 751–777.

[27] L. van den Dries, *A generalization of the Tarski-Seidenberg theorem, and some nondefinability results*, Bull. Amer. Math. Soc. (N.S.) **15** (1986), no. 2, 189–193.

[28] L. van den Dries and C. Miller, *On the real exponential field with restricted analytic functions*, Israel J. Math. **85** (1994), no. 1-3, 19–56.
[29] C. Miller, *Expansions of the real field with power functions*, Ann. Pure Appl. Logic **68** (1994), no. 1, 79–94.
[30] L. van den Dries, A. Macintyre, and D. Marker, *Logarithmic-exponential power series*, J. Lond. Math. Soc. **56** (1997), no. 3, 417–434.
[31] P. Speissegger, *Pfaffian sets and o-minimality*, Lecture notes on o-minimal structures and real analytic geometry, Springer, 2012, pp. 179–217.
[32] L. van den Dries and P. Speissegger, *The real field with convergent generalized power series*, Trans. Amer. Math. Soc. **350** (1998), no. 11, 4377–4421.
[33] ———, *The field of reals with multisummable series and the exponential function*, Proc. Lond. Math. Soc. **81** (2000), no. 3, 513–565.
[34] C. Miller and P. Speissegger, *Expansions of the real field by canonical products*, Canad. Math. Bull. **63** (2020), no. 3, 506–521.
[35] A. Greenbaum, *Iterative methods for solving linear systems*, Frontiers in Applied Mathematics, SIAM, 1997.
[36] M. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Stand. **49** (1952), no. 6, 409–436.
[37] Z. Strakoš and P. Tichý, *Error estimation in preconditioned conjugate gradients*, BIT **45** (2005), no. 4, 789–817.
[38] ———, *On error estimation in the conjugate gradient method and why it works in finite precision computations*, Electron. Trans. Numer. Anal. **13** (2002), 56–80.
[39] G. Meurant, *The Lanczos and conjugate gradient algorithms: from theory to finite precision computations*, Society for Industrial and Applied Mathematics, 2006.
[40] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl. **113** (1989), 7–63.
[41] A. Greenbaum and Z. Strakoš, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl. **13** (1992), no. 1, 121–137.
[42] M. Nikolova, *Markovian reconstruction using a GNC approach*, IEEE Trans. Image Process. **8** (1999), no. 9, 1204–1220.
[43] I. Borg, P. Groenen, and P. Mair, *Applied multidimensional scaling and unfolding*, Springer, 2018.
[44] B. Bollobás, *Modern graph theory*, Springer, 1998.
[45] J. de Leeuw, *Differentiability of Kruskal's stress at a local minimum*, Psychometrika **49** (1984), no. 1, 111–113.
[46] ———, *Fitting distances by least squares*, Technical Report 130, Interdivisional Program in Statistics, UCLA, 1993.
[47] ———, *Applications of convex analysis to multidimensional scaling*, Proc. European Meeting of Statisticians, pp. 133–146, Grenoble, France, Sept. 1976.
[48] ———, *Convergence of the majorization method for multidimensional scaling*, J. Classif. **5** (1988), no. 2, 163–180.
[49] P. Groenen and M. van de Velden, *Multidimensional scaling by majorization: a review*, J. Stat. Softw. **73** (2016), no. 8.
[50] M. Bertero and P. Boccacci, *Introduction to inverse problems in imaging*, CRC Press, 1998.
[51] A. Ribes and F. Schmitt, *Linear inverse problems in imaging*, IEEE Signal Processing Mag. **25** (2008), no. 4, 84–99.
[52] P. Hansen, *Discrete inverse problems: insight and algorithms*, SIAM, 2010.
[53] M. Robini and Y. Zhu, *Generic half-quadratic optimization for image reconstruction*, SIAM J. Imaging Sci. **8** (2015), no. 3, 1752–1797.
[54] P. Liu, L. Xiao, and J. Zhang, *A fast higher degree total variation minimization method for image restoration*, Int. J. Comput. Math. **93** (2016), no. 8, 1383–1404.
[55] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math. **57** (2004), no. 11, 1413–1457.
[56] Y.-R. Li, L. Shen, D.-Q. Dai, and B. Suter, *Framelet algorithms for de-blurring images corrupted by impulse plus Gaussian noise*, IEEE Trans. Image Process. **20** (2011), no. 7, 1822–1837.

[57] S. Ravishankar and Y. Bresler, *Learning sparsifying transforms*, IEEE Trans. Signal Processing **61** (2013), no. 5, 1072–1086.
[58] S. Li, *Markov random field modeling in image analysis*, Springer, 2001.
[59] I. Tollis, G. Di Battista, P. Eades, and R. Tamassia, *Graph drawing: algorithms for the visualization of graphs*, Prentice Hall, 1999.
[60] U. Brandes, *Drawing on physical analogies*, Lecture Notes in Comput. Sci. **2025** (2001), 71–86.
[61] T. Kamada and S. Kawai, *An algorithm for drawing general undirected graphs*, Inform. Process. Lett. **31** (1989), no. 1, 7–15.
[62] E. Gansner, Y. Koren, and S. North, *Graph drawing by stress majorization*, Proc. 12th Int. Symp. Graph Drawing, Lecture Notes in Comput. Sci. **3383** (2005), 239–250.
[63] T. Davis and Y. Hu, *The University of Florida sparse matrix collection*, ACM Trans. Math. Software **38** (2011), no. 1.
[64] K. Hämäläinen, L. Harhanen, A. Kallonen, A. Kujanpää, E. Niemi, and S. Siltanen, *Tomographic X-ray data of a walnut*, arXiv:1502.04064, 2015.
[65] R. Acar and C. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems **10** (1994), no. 6, 1217–1229.
[66] M. Black, G. Sapiro, D. Marimont, and D. Heeger, *Robust anisotropic diffusion*, IEEE Trans. Image Process. **7** (1998), no. 3, 421–432.
[67] J.-F. Cai, R. Chan, L. Shen, and Z. Shen, *Convergence analysis of tight framelet approach for missing data recovery*, Adv. Comput. Math. **31** (2009), no. 1, 87–113.
[68] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, *Image quality assessment: from error visibility to structural similarity*, IEEE Trans. Image Process. **13** (2004), no. 4, 600–612.
[69] R. S. Dembo and T. Steihaug, *Truncated-Newton algorithms for large-scale unconstrained optimization*, Math. Program. **26** (1983), 190–212.
[70] J. C. Gilbert and J. Nocedal, *Global convergence properties of conjugate gradient methods for optimization*, SIAM J. Optim. **2** (1992), no. 1, 21–42.
[71] J. Nocedal and S. Wright, *Numerical optimization*, 2nd ed., Springer Series in Operations Research and Financial Engineering, 2006.
[72] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Math. Program. **45** (1989), 503–528.

M. Robini and Y. Zhu are with the CREATIS laboratory (CNRS research unit UMR5220 and INSERM research unit U1294), INSA Lyon, France.

Email address: marc.robini@creatis.insa-lyon.fr, zhu@creatis.insa-lyon.fr

L. wang is with the Key Laboratory of Intelligent Medical Image Analysis and Precise Diagnosis, Guizhou University, Guiyang, China.

Email address: lhwang2@gzu.edu.cn

Revision Report — JOGO-D-22-00253

The Appeals of Quadratic Majorization-Minimization

Marc C. Robini, Lihui Wang, and Yuemin Zhu

We gratefully thank the referee for evaluating our work and we apologize for our late answer (the referee's comments were made available to us on June 2023, for reasons independent from the editor,).

All recommendations have been carefully taken into account. Each of our answers is preceded by the associated original comment typeset in a *blue slanted font*; highlights and changes in the manuscript are displayed in red. Please note that this report has its own reference section.

1. *In section 13, the authors point out that the algorithm can be applied to problems such as multidimensional scaling and regularized linear inversion, and give some of the details. Why the authors do not give the corresponding numerical experiments in the last section?*

Our numerical experiments are on graph layout and X-ray tomography, which are instances of multidimensional scaling and regularized linear inversion, respectively. To avoid confusion, we make it clear in the introduction as well as in the experimental section:

- In the second-to-last paragraph of the introduction (page 3):

  "Section 14 concludes the paper with numerical experiments on graph layout (an instance of multidimensional scaling) and X-ray tomography (an instance of regularized inversion)."

- At the beginning of Section 14 (page 31):

  "In this section we present experiments on graph layout and X-ray tomography as instances of multidimensional scaling and regularized inversion, respectively."

- In the second paragraph of Section 14.1 (page 31), where we explain how the energy function for graph layout is a special case of the stress function for multidimensional scaling:

  "The Kamada-Kawai energy is a special case of the raw stress function (13.1) in which $q = 2$ and the matrix $[c_{ij}]$ is proportional to $[1/d_{ij}^2]$, that is,

  $$F(\boldsymbol{X}) \propto \sum_{i,j=1}^{m} \big(1 - \big\|\boldsymbol{X}^{(i)} - \boldsymbol{X}^{(j)}\big\|/d_{ij}\big)^2, \tag{14.2}$$

  where $\boldsymbol{X}^{(i)}$ and $\boldsymbol{X}^{(j)}$ contain the two-dimensional cartesian coordinates of v_i and v_j, respectively. The corresponding objective f and its surrogate are defined by (13.15) and (13.16)."

- In the second paragraph of Section 14.2 (page 40), where we explain how the objective function for X-ray tomography is a special case of that of regularized linear inversion:

  "We look for solutions that minimize

  $$f(\boldsymbol{x}) := \|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{d}\|^2 + \lambda \sum_{j=1}^{l} \theta\big(\|\boldsymbol{R}_j\boldsymbol{x}\|/\delta\big), \tag{14.12}$$

where $\lambda > 0$ controls the regularization strength and $\delta > 0$ adjusts the scale of the operator $\{\boldsymbol{R}_j\}_j$. This objective has the form (13.22) with $M = m + l$ and

$$(\boldsymbol{A}_i, \boldsymbol{a}_i, \theta_i(t)) = \begin{cases} (\boldsymbol{D}(i,:), [\boldsymbol{d}]_i, t^2) & \text{if } i \leqslant m, \\ (\boldsymbol{R}_{i-m}, \boldsymbol{0}, \lambda\theta(t/\delta)) & \text{otherwise."} \end{cases} \tag{14.13}$$

2. *Four matrices from the SuiteSparse Matrix Collection are used for the numerical experiments in Section 14.1. What are the reasons for choosing these four matrices? The authors give numerical results for only two matrices in this paper. I suggest that the authors put the numerical results for the remaining two matrices and add more examples to make the numerical experiments more convincing.*

We selected the matrices `1138_bus`, `tuma2`, `filter2D` and `rajat19` because we find them representative of the layouts obtained by minimizing the multidimensional scaling stress function. To reinforce our results, we now provide additional layout examples as well as the numerical results for `filter2D` and `rajat19`. More specifically, we made the following additions:

- Figure 4 (page 34) shows the layouts of the graphs `dwt_1242`, `airfoil1_dual` and `uk` (also from the SuiteSparse Matrix Collection) produced by the PCG-QMM algorithm.
- Table 1 (page 32) gives the numbers of vertices and edges of the graphs considered in the experiments, together with the spectral condition number of their corresponding inner system matrix.
- Table 2 (page 35) and Figure 5 (page 36) are augmented with the numerical results associated with `filter2D` and `rajat19`, which confirm the effects of the accuracy of the PCG solver observed for `1138_bus` and `tuma2`.
- Figure 6 (page 36) and Figure 8 (page 39) plot the gradient norm, the normalized objective and the distance to the limit versus the number of outer iterations for `filter2D` and `rajat19`. The observations are similar to those for `1138_bus` and `tuma2`.
- Table 3 (page 40) and Table 5 (page 41) are augmented with the numerical results associated with `filter2D` and `rajat19` for further illustrating the behavior of PCG-QMM in the long run and the reliability of the PCG stopping criterion.

3. *It is suggested that the authors illustrate the effectiveness of the algorithm by comparing it with some other methods. For example, in Section 14.2, a comparison with Gondzio's method [GLLA+22] and the Joint Total Variation (JTV) method mentioned in his paper [TMSK20] can be considered.*

The methods described in [GLLA+22] and [TMSK20] are specific to multi-energy X-ray tomography, which is beyond the scope of our experiments. However, the JTV regularizer considered in [TMSK20] fits into our framework for regularized linear inversion (the inner product regularizer proposed in [GLLA+22] does not, because the reversal matrix L is not positive). In [TMSK20] the optimization problem is solved using a Polak-Ribière conjugate gradient method (CG-PR); so to meet the referee's requirement, we compare the PCGLS-QMM algorithm with the CG-PR+ algorithm [GN92] as well as with two other nonlinear CG-related algorithms: truncated newton (TN) [DS83] and L-BFGS [NW06]. The comparison results are given in Section 14.2.5 (page 46) whose content is reproduced below.

> "Finally, we compare PCGLS-QMM with three nonlinear CG-related methods: truncated Newton (TN) [DS83], Polak-Ribière conjugate gradient (CG-PR+) [GN92], and L-BFGS [NW06]. These three algorithms are major tools for large-scale optimization, but they do

not come with as strong convergence guarantees as PCGLS-QMM. The limited memory parameter of L-BFGS is set to the recommended value of 5 [LN89], and we consider PCGLS-QMM in fast and normal settings (namely, $(k, \gamma) = (1, 0.1)$ and $(5, 10^{-2})$, respectively). For an outer termination tolerance of 10^{-5}, TN, CG-PR+ and L-BFGS all produce similar reconstructions to PCGLS-QMM: the objective, PSNR and SD are the same to four significant digits in both the convex and nonconvex cases. However, as Table 8 shows, there are significant differences in running time: PCGLS-QMM is 7 to 14 times faster than TN and CG-PR+, and up to twice as fast as L-BFGS. Furthermore, in the long run, the accuracy of TN, CG-PR+ and L-BFGS plateaus prematurely compared to PCGLS-QMM. This is illustrated in Table 9, which gives the number of iterations to and the value of the minimum gradient norm: the gradient norm of the iterates of TN, CG-PR+ and L-BFGS plateaus above 10^{-6} after 400–700 iterations, while PCGLS-QMM plateaus below 1.5×10^{-13} after 1400–4000 iterations."

4. *In general, convex problems are easier to solve than non-convex problems. Why do convex problems take more iterations than non-convex problems in 14.2.4? Besides, I think it is not convincing to illustrate the robustness to the parameters of the PCG solver by just one tomography problem.*

First we note that solving the convex problem is faster in terms of both the running time and the number of outer iterations. This is mentioned in Section 14.2.2 (page 42) :

> "In the convex case, the running time ranges from about 2 minutes for $(k, \gamma) = (1, 0.1)$ to about 10 minutes for $(k, \gamma) = (32, 10^{-8})$, and the number N_{MM} of outer iterations is between 220 and 225. In the nonconvex case, the running time ranges from 15 to 52 minutes, and N_{MM} is between 360 and 400 when $k \geqslant 4$ and goes up to 660 for $(k, \gamma) = (1, 0.1)$."

However, we agree that the results in Section 14.2.4 need to be clarified; so we added the following paragraph (page 44) to explain why the stopping delay k for matching the target contraction γ is larger in the convex case than in the nonconvex one :

> "We notice that the minimum suitable value of k for matching γ is larger in the convex case. This indicates that, along the iterate trajectories, the spectra of the surrogates (that is, the distributions of the eigenvalues of the inner system matrices) are more spread out for the convex objective than for the nonconvex one. Indeed, we have seen in Section 12.2 that the larger k, the better the approximation (12.10) of the inner error energy norm, and hence the closer the inner contraction to γ. At the same time, the error bound (12.13) not only shows that the quality of this approximation depends on the spectrum of the inner system matrix, but also gives insight into what qualifies as a favorable spectrum, namely tightly clustered eigenvalues away from the origin, and what constitutes an unfavorable one, namely widely spread out eigenvalues (see [Gre97, Section 3.1]). Therefore, the more scattered the eigenvalues, the larger k must be for matching γ."

Finally, we now also illustrate the robustness to the parameters of the PCG solver in the graph layout experiments (Section 14.1.3, page 37) :

> "We set the outer termination tolerance ϵ to 10^{-6} and look at the behavior of the PCG-QMM algorithm for a stopping delay ranging from 1 to 32 and a contraction number between 10^{-8} and 10^{-1}. Table 4 gives the ranges of (i) the relative objective difference between the pratical

and double-limit solutions $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{x}}^{**}$, (ii) the Tucker distance between $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{x}}^{**}$, (iii) the running time, and (iv) the number of outer iterations. We see that the practical solutions are very close to the double limit solutions, and therefore not sensitive to the parameters of the PCG solver. Furthermore, the maximum-to-minimum ratios of the running time and of the number of outer iterations (less than 4 and 2, respectively) are small relative to the ranges of k and γ. The running time is maximal for $(k, \gamma) = (32, 10^{-8})$ and minimal or close to minimal for $(k, \gamma) = (2, 10^{-2})$, and the number of outer iterations is maximal for $(k, \gamma) = (1, 10^{-1})$ and stabilizes as k increases and/or γ decreases."

References

[DS83] R. S. Dembo and T. Steihaug. Truncated-Newton algorithms for large-scale unconstrained optimization. *Math. Program.*, 26:190–212, 1983.

[GLLA+22] J. Gondzio, M. Lassas, S.-M. Latva-Äijö, S. Siltanen, and F. Zanetti. Material-separating regularizer for multi-energy x-ray tomography. *Inverse Problems*, 28(2), 2022.

[GN92] J. C. Gilbert and J. Nocedal. Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Optim.*, 2(1):21–42, 1992.

[Gre97] A. Greenbaum. *Iterative methods for solving linear systems.* Frontiers in Applied Mathematics. SIAM, 1997.

[LN89] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45:503–528, 1989.

[NW06] J. Nocedal and S. Wright. *Numerical optimization.* Springer Series in Operations Research and Financial Engineering. 2nd edition, 2006.

[TMSK20] J. Toivanen, A. Meaney, S. Siltanen, and V. Kolehmainen. Joint reconstruction in low dose multi-energy CT. *Inverse Probl. Imaging*, 14(4):607–629, 2020.