Properties of the delayed weighted gradient method

Roberto Andreani*   Marcos Raydan†

March 2, 2020

Abstract

The delayed weighted gradient method, recently introduced in [13], is a low-cost gradient-type method that exhibits a surprisingly and perhaps unexpected fast convergence behavior that competes favorably with the well-known conjugate gradient method for the minimization of convex quadratic functions. In this work, we establish several orthogonality properties that add understanding to the practical behavior of the method, including its finite termination. We show that if the $n \times n$ real Hessian matrix of the quadratic function has only $p < n$ distinct eigenvalues, then the method terminates in p iterations. We also establish an optimality condition, concerning the gradient norm, that motivates the use of this novel scheme when low precision is required for the minimization of non-quadratic functions.

Keywords: Gradient methods, conjugate gradient methods, smoothing techniques, finite termination, Krylov subspace methods.

1 Introduction

Recently [13], Oviedo proposed a low-cost method for the minimization of large-scale convex quadratic functions, which is based on a smoothing technique combined with a one-step delayed gradient method. For gradient-type methods, smoothing techniques were previously developed [1, 11], as well as delayed schemes [7, 12]. A skillful combination of these independent ideas produces the so-called delayed weighted gradient method (DWGM), which exhibits an impressive fast convergence behavior that compares favorable with the conjugate gradient (CG) method [13]. Moreover, we have observed that in exact arithmetic DWGM also exhibits finite termination, which is a well-known property of the CG method, as well as the Conjugate Residual (CR) method, in the convex quadratic case [14]. Nevertheless, from the same initial guess, DWGM produces a different sequence of

---

*Department of Applied Mathematics, IMECC-UNICAMP, University of Campinas, Rua Sérgio Buarque de Holanda, 651 Cidade Universitária "Zeferino Vaz", Distrito Barão Geraldo, 13083-859 Campinas SP, Brazil (andreani@unicamp.br).

†Centro de Matemática e Aplicações (CMA), FCT, UNL, 2829-516 Caparica, Portugal (m.raydan@fct.unl.pt).

iterates to converge (or to terminate) at the same solution obtained by the CG or the CR method (see Table 1 and Figure 1 below). In this work we will establish several properties of the DWGM method, including its surprising finite termination for convex quadratics.

The rest of this document is organized as follows. In Section 2, we briefly describe the DWGM algorithm and list the convergence properties established in [13]. In Section 3, we describe and establish the additional properties concerning the orthogonality relationships that exist among the involved sequence of vectors produced by the DWGM algorithm, and we also establish the key results concerning finite termination and optimality of the gradient norm. In Section 4, we present some final remarks and perspectives.

2   DWGM algorithm

Let us consider the minimization of the strictly convex quadratic function

$$f(x) = \frac{1}{2}x^T A x - b^T x, \tag{2.1}$$

where $b \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ is a symmetric and positive definite (SPD) matrix. Since A is SPD and the gradient $g(x) \equiv \nabla f(x) = Ax - b$, then the global solution of (2.1) is the unique solution $A^{-1}b$ of the linear system $Ax = b$. For n large, many low-cost iterative methods have been proposed and analyzed, and in particular some of the so-called gradient-type methods have shown to be very competitive since they show an impressive fast linear convergence; see, e.g., [2, 3, 4, 5, 6, 9]. They can all be seen as improved extensions of the classical steepest descent method. From a starting point $x_0 \in \mathbb{R}^n$, the well-known steepest descent (or gradient) method is given by the iteration

$$x_{k+1} = x_k - \lambda_k g_k,$$

where $g_k = g(x_k)$ and λ_k is the minimizer of $f(x_k - \lambda g_k)$. Therefore, if $g_k \neq 0$,

$$\lambda_k = \frac{g_k^T g_k}{g_k^T A g_k}. \tag{2.2}$$

This classical low-cost method is globally convergent but its rate of convergence is very slow in most practical cases.

We focus our attention on the DWGM method, that can be viewed as a special gradient-type method. Following the development in [13], the DWGM algorithm is now presented.

2

Algorithm 1 DWGM
___
Require: $A \in \mathbb{R}^{n \times n}$ SPD, $x_0 \in \mathbb{R}^n$, $x_{-1} = x_0$, $g_0 = g(x_0)$, $g_{-1} = g_0$, $\varepsilon > 0$.

1: $k \leftarrow 0$

2: **while** $\|g_k\|_2 > \varepsilon$ **do**

3: $\quad w_k = Ag_k$

4: $\quad \alpha_k = g_k^T w_k / w_k^T w_k$

5: $\quad y_k = x_k - \alpha_k g_k$

6: $\quad r_k = g_k - \alpha_k w_k$

7: $\quad \beta_k = g_{k-1}^T (g_{k-1} - r_k) / \|g_{k-1} - r_k\|_2^2$

8: $\quad x_{k+1} = x_{k-1} + \beta_k (y_k - x_{k-1})$

9: $\quad g_{k+1} = g_{k-1} + \beta_k (r_k - g_{k-1})$

10: $\quad k \leftarrow k + 1$

11: **end while**
___

Notice that the iterations in Algorithm DWGM are stopped when the 2-norm of the gradient at x_k is less than or equal to a preestablished small tolerance $\varepsilon > 0$. Note also that in addition to the sequence of iterates $\{x_k\}$, the algorithm DWGM generates an auxiliary sequence $\{y_k\}$, and two associated gradient-type sequences given by

$$g_k = Ax_k - b \quad \text{and} \quad r_k = Ay_k - b. \tag{2.3}$$

Finally, we note that at every k, the scalar β_k is chosen at Step 7 to guarantee that $g_{k+1}^T (r_k - g_{k-1}) = 0$, and the scalar α_k is chosen at Step 4 to minimize the gradient norm along the negative gradient direction, as explained in [13, p. 731]; see also [5].

In [13], it is established that $\beta_k \geq 0$ for all k, that the sequence $\{\|g_k\|_2\}$ is monotonically decreasing, and moreover that the sequence $\{g_k\}$ converges to zero q-linearly when k goes to infinity, which implies that the sequence $\{x_k\}$ converges to the unique global minimizer of $f(x)$. Furthermore, from the proof of Lemma 1 in [13], if $g_k \neq 0$ we notice that

$$\|r_k\|_2^2 = \|g_k\|_2^2 - \alpha_k g_k^T A g_k,$$

and also that

$$\|g_{k+1}\|_2 \leq \|r_k\|_2 < \|g_k\|_2 \leq \|r_{k-1}\|_2. \tag{2.4}$$

Since α_k can be seen as a Rayleigh quotient of A evaluated at the vector $A^{1/2} g_k$, then for all k

$$0 < \lambda_{\min}(A) \leq \alpha_k \leq \lambda_{\max}(A) < \infty, \tag{2.5}$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the smallest and largest eigenvalues of A, respectively.

3

3 Additional properties of DWGM

Let us recall that, in exact arithmetic, the A-orthogonality of the set of all search directions plays a fundamental role in the finite termination of the CG method. Similarly, the A-orthogonality of the set of all residual vectors plays a fundamental role in the finite termination of the CR method see, e.g., [14]. We note that a similar A-orthogonality is not imposed on any of the vector sets used by DWGM, however it also exhibits finite termination for convex quadratic functions, as can be observed in Table 1 and Figure 1. In order to establish this important fact we need a few preliminary results, which include a few unexpected A-orthogonality relationships.

Lemma 1. *In Algorithm DWGM*

(a) $\beta_0 = 1$, and hence $x_1 = y_0$ and $g_1 = r_0$.

The following equalities hold for all $k \geq 0$

(b) $r_k^T A g_k = 0$.

(c) $r_k^T r_k = r_k^T g_k$.

(d) $g_{k+1}^T (r_k - g_{k-1}) = 0$.

Proof. (a) Since $g_{-1} = g_0$, using the definition of β_k at Step 7 and α_k at Step 4, combined with the equality at Step 6, we obtain

$$\beta_0 = g_0^T (g_0 - r_0)/\|g_0 - r_0\|_2 = \alpha_0 (g_0^T w_0)/\alpha_0^2 \|w_0\|_2^2 = g_0^T A g_0/(\alpha_0 \, g_0^T A^2 g_0) = 1.$$

Now, using $\beta_0 = 1$ in Steps 8 and 9, it follows that $x_1 = y_0$ and $g_1 = r_0$.
(b) Combining Steps 6, 4, and 3, we have

$$r_k^T A g_k = (g_k - \alpha_k A g_k)^T A g_k = g_k^T A g_k - \alpha_k (g_k^T A^2 g_k) = 0.$$

(c) Using Step 6 and (b) we obtain

$$r_k^T r_k - r_k^T g_k = r_k^T (r_k - g_k) = -\alpha_k (r_k^T A g_k) = 0.$$

(d) Follows from the definition of β_k at Step 7 and simple algebraic manipulations. □

We note that the choice of α_k accounts for $r_k^T A g_k = 0$ for all $k \geq 0$, which is an A-orthogonality result that will play a key role in the rest of this section.

Lemma 2. *Algorithm DWGM generates the sequence of gradient vectors $\{g_k\}$ such that for all $k \geq 1$,*

$$g_k^T A g_{k-1} = 0.$$

Proof. Using (a) and (b) from Lemma 1, it follows that $g_1 = r_0$ and $g_1^T A g_0 = r_0^T A g_0 = 0$, and the result holds for $k = 1$. Let us now assume, by induction on k, that

4

Table 1: Gradient norm at each iteration (Iter i) of DWGM, CG, and CR, from the same initial point, for the minimization of a strictly convex quadratic function when $n = 5$, and the eigenvalues of A are given by $\lambda_i = i$, for $1 \leq i \leq 5$.

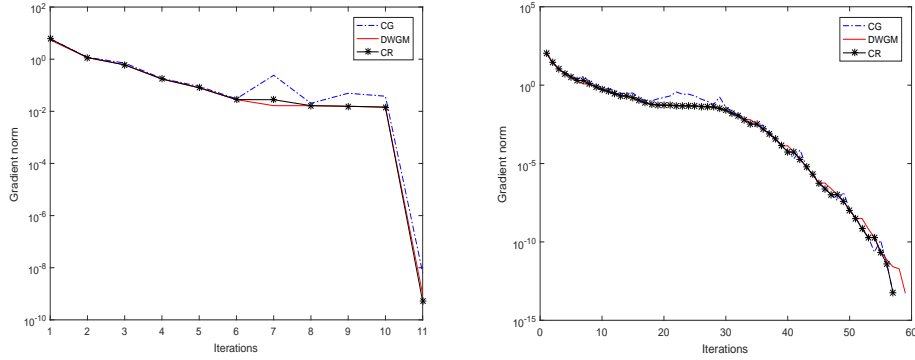| Method	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5
CG	1.3347	0.6458	0.2418	0.0368	0.0
CR	1.2934	0.5778	0.2231	0.0363	0.0
DWGM	0.6679	0.2259	0.0442	0.0064	0.0

Figure 1: Convergence history of DWGM, CG, and CR, from the same initial point, for the minimization of a strictly convex quadratic function. On the left, $n = 11$ and we use the following distribution of eigenvalues of A: $\lambda_1 = 0.1$, $\lambda_i = i$ for $2 \leq i \leq 10$, and $\lambda_{11} = 1000$. On the right, $n = 60$ and we use the following distribution of eigenvalues of A: $\lambda_1 = 0.1$, $\lambda_i = i$ for $2 \leq i \leq 59$, and $\lambda_{10} = 6000$.

$g_k^T A g_{k-1} = 0$ up to $k = \hat{k} \geq 2$, and consider the next iteration. Hence, we need to show that $g_{\hat{k}+1}^T A g_{\hat{k}} = 0$. Using Step 9, (b) from Lemma 1, and noticing that by the inductive hypothesis $g_{\hat{k}-1}^T A g_{\hat{k}} = 0$, we have

$$g_{\hat{k}+1}^T A g_{\hat{k}} = (g_{\hat{k}-1} + \beta_{\hat{k}}(r_{\hat{k}} - g_{\hat{k}-1}))^T A g_{\hat{k}} = (1 - \beta_{\hat{k}})g_{\hat{k}-1}^T A g_{\hat{k}} + \beta_{\hat{k}} r_{\hat{k}}^T A g_{\hat{k}} = 0,$$

and the result is established. □

Lemma 3. *In Algorithm DWGM the following statements hold for all $k \geq 1$*

(a) $g_k^T r_{k-1} = g_k^T g_{k-1}$.

(b) $g_{k-1}^T r_k = g_{k-1}^T g_k$.

(c) $g_k^T r_{k-1} = g_{k-1}^T r_k$.

(d) $\|g_{k-1} - r_k\|_2^2 = \|g_{k-1} - g_k\|_2^2 - [(g_k^T A g_k)^2 / g_k^T A^2 g_k]$.

(e) $g_{k+1}^T (g_k - g_{k-1}) = 0$.

(f) $\beta_k > 0$.

Proof. (a) Using Steps 4 and 6 and Lemma 2, we have

$$g_k^T (r_{k-1} - g_{k-1}) = -\alpha_{k-1} g_k^T A g_{k-1} = 0.$$

(b) Similar to the proof of (a),

$$g_{k-1}^T (r_k - g_k) = -\alpha_k g_{k-1}^T A g_k = 0.$$

(c) The result is obtained directly from (a) and (b).

(d) Using Step 4 and 6 of the DWGM algorithm and Lemma 2, we have

$$
\begin{aligned}
\|g_{k-1} - r_k\|_2^2 &= \|g_{k-1} - g_k + \alpha_k A g_k\|_2^2 \\
&= \|g_{k-1} - g_k\|_2^2 + 2(g_{k-1} - g_k)^T \alpha_k A g_k + \alpha_k^2 g_k^T A^2 g_k \\
&= \|g_{k-1} - g_k\|_2^2 - 2\alpha_k g_k^T A g_k + \alpha_k^2 g_k^T A^2 g_k \\
&= \|g_{k-1} - g_k\|_2^2 - [(g_k^T A g_k)^2 / g_k^T A^2 g_k].
\end{aligned}
$$

(e) Combining Step 6 of the DWGM algorithm, (d) from Lemma 1, and Lemma 2, it follows that

$$g_{k+1}^T (g_k - g_{k-1}) = g_{k+1}^T (r_k + \alpha_k A g_k - g_{k-1}) = g_{k+1}^T (r_k - g_{k-1}) + \alpha_k g_{k+1}^T A g_k = 0.$$

(f) By Cauchy-Schwarz inequality and (2.4), we obtain

$$g_{k-1}^T g_k \leq \|g_{k-1}\|_2 \|g_k\|_2 < \|g_{k-1}\|_2^2 = g_{k-1}^T g_{k-1},$$

and hence, using (b), we have that $g_{k-1}^T (g_{k-1} - r_k) = g_{k-1}^T (g_{k-1} - g_k) > 0$. Therefore, the numerator at Step 7 of the DWGM algorithm is positive and $\beta_k > 0$. □

Lemma 4. *In Algorithm DWGM the following statements hold for all $k \geq 1$*

(a) $g_{k+1}^T g_{k+1} = g_{k+1}^T g_k = g_{k+1}^T g_{k-1}$.

(b) $\|g_{k-1} - g_k\|_2^2 = g_{k-1}^T (g_{k-1} - g_k)$

(c) $\beta_k = g_{k-1}^T (g_{k-1} - g_k) / \left(\|g_{k-1} - g_k\|_2^2 - [(g_k^T A g_k)^2 / g_k^T A^2 g_k] \right) > 1$.

Proof. (a) Combining Steps 3, 6, and 9 of the DWGM algorithm, we get

$$g_{k+1}^T g_{k+1} = g_{k+1}^T (g_{k-1} + \beta_k (g_k - g_{k-1} - \alpha_k A g_k)) = g_{k+1}^T g_{k-1} + \beta_k g_{k+1}^T (g_k - g_{k-1}) - \alpha_k g_{k+1}^T A g_k.$$

From (e) in Lemma 3, $g_{k+1}^T(g_k - g_{k-1}) = 0$, and from Lemma 2 we have that $g_{k+1}^T Ag_k = 0$, and so $g_{k+1}^T g_{k+1} = g_{k+1}^T g_{k-1}$. Now, using again (e) in Lemma 3, we obtain that $g_{k+1}^T g_{k-1} = g_{k+1}^T g_k$.

(b) From (a) we obtain $g_k^T(g_k - g_{k-1}) = 0$, and hence

$$0 = (g_k - g_{k-1} + g_{k-1})^T(g_k - g_{k-1}) = \|g_k - g_{k-1}\|_2^2 - g_{k-1}^T(g_{k-1} - g_k),$$

which implies that $\|g_{k-1} - g_k\|_2^2 = g_{k-1}^T(g_{k-1} - g_k)$.

(c) Combining Step 7 of the DWGM algorithm with (b) and (d) in Lemma 3 we get

$$\beta_k = \frac{g_{k-1}^T(g_{k-1} - r_k)}{\|g_{k-1} - r_k\|_2^2} = \frac{g_{k-1}^T(g_{k-1} - g_k)}{\left(\|g_{k-1} - g_k\|_2^2 - [(g_k^T Ag_k)^2/g_k^T A^2 g_k]\right)}. \tag{3.1}$$

Now, from (f) in Lemma 3 we have that β_k is strictly positive and so

$$g_{k-1}^T(g_{k-1} - g_k) = g_{k-1}^T(g_{k-1} - r_k) > 0,$$

and hence the denominator in (3.1) must be also positive. As a consequence

$$0 < (g_k^T Ag_k)^2/g_k^T A^2 g_k < \|g_{k-1} - g_k\|_2^2,$$

and we conclude using (b) that in (3.1) the numerator is strictly bigger than the denominator and both are positive, which establishes that $\beta_k > 1$. □

In what follows we will establish some key A-orthogonality results, which will be obtained simultaneously using an inductive argument.

Theorem 5. *Algorithm DWGM generates the sequences $\{g_k\}$ and $\{r_k\}$ such that*

(i) For $k \geq 2$, $g_k^T Ag_j = 0$ for all $-1 \leq j \leq k - 2$.

(ii) For $k \geq 2$, $r_k^T Ag_j = 0$ for all $-1 \leq j \leq k - 2$.

Proof. Concerning (i), since $g_0 = g_{-1}$, $g_1 = r_0$ and $\alpha_0 > 0$ (see (2.5)), using (e) from Lemma 3 and Steps 4 and 6 of the DWGM algorithm we obtain

$$g_2^T Ag_{-1} = g_2^T Ag_0 = g_2^T(g_0 - r_0)/\alpha_0 = g_2^T(g_0 - g_1)/\alpha_0 = 0, \tag{3.2}$$

and the result is obtained for $k = 2$. Concerning (ii), since $\alpha_k > 0$ for all k (see (2.5)), using (e) in Lemma 3, (a) in Lemma 1, Step 6 (twice), Lemma 2, and (3.2), it follows that

$$
\begin{aligned}
r_2^T Ag_0 &= \frac{1}{\alpha_0} r_2^T(g_0 - r_0) = -\frac{1}{\alpha_0} r_2^T(g_1 - g_0) + 0 = -\frac{1}{\alpha_0} r_2^T(g_1 - g_0) + \frac{1}{\alpha_0} g_2^T(g_1 - g_0) \\
&= \frac{1}{\alpha_0}(g_2 - r_2)^T(g_1 - g_0) = \frac{\alpha_2}{\alpha_0} g_2^T A(g_1 - g_0) = \frac{\alpha_2}{\alpha_0}[g_2^T Ag_1 - g_2^T Ag_{-1}] = 0.
\end{aligned}
$$

7

Since $r_2^T A g_{-1} = r_2^T A g_0 = 0$, the result is established for $k = 2$.

Let us now assume, by induction on k, that (i) and (ii) hold up to $k = \hat{k} \geq 3$, and consider the next iteration. Hence, we need to show that $g_{\hat{k}+1}^T A g_j = 0$, and also that $r_{\hat{k}+1}^T A g_j = 0$, for all $-1 \leq j \leq \hat{k} - 1$.

For $-1 \leq j \leq \hat{k} - 2$, using Lemma 2, Step 9 of the DWGM algorithm, and the inductive hypothesis associated with (i) and (ii), we have that

$$g_{\hat{k}+1}^T A g_j = (g_{\hat{k}-1} + \beta_{\hat{k}}(r_{\hat{k}} - g_{\hat{k}-1}))^T A g_j = (1 - \beta_{\hat{k}}) g_{\hat{k}-1}^T A g_j + \beta_{\hat{k}} r_{\hat{k}}^T A g_j = 0.$$

For $j = \hat{k} - 1$, using Step 6, adding and subtracting $g_{\hat{k}-2}$, and then using the fact that $r_{\hat{k}-1} - g_{\hat{k}-2} = (g_{\hat{k}} - g_{\hat{k}-2})/\beta_{\hat{k}-1}$ (from Step 9 of DWGM algorithm), we get

$$
\begin{aligned}
g_{\hat{k}+1}^T A g_{\hat{k}-1} &= \frac{1}{\alpha_{\hat{k}-1}} g_{\hat{k}+1}^T (g_{\hat{k}-1} - r_{\hat{k}-1}) = -\frac{1}{\alpha_{\hat{k}-1}} g_{\hat{k}+1}^T (r_{\hat{k}-1} - g_{\hat{k}-2} + g_{\hat{k}-2} - g_{\hat{k}-1}) \\
&= -\frac{1}{\alpha_{\hat{k}-1}} \left[\frac{1}{\beta_{\hat{k}-1}} g_{\hat{k}+1}^T (g_{\hat{k}} - g_{\hat{k}-2}) + g_{\hat{k}+1}^T (g_{\hat{k}-2} - g_{\hat{k}-1}) \right].
\end{aligned}
$$

Adding and subtracting $g_{\hat{k}+1}^T g_{\hat{k}-1}$, using (e) from Lemma 3, and Step 9, we obtain

$$
\begin{aligned}
g_{\hat{k}+1}^T A g_{\hat{k}-1} &= -\frac{1}{\alpha_{\hat{k}-1}} \left[\frac{g_{\hat{k}+1}^T (g_{\hat{k}} - g_{\hat{k}-1}) + g_{\hat{k}+1}^T (g_{\hat{k}-1} - g_{\hat{k}-2})}{\beta_{\hat{k}-1}} + g_{\hat{k}+1}^T (g_{\hat{k}-2} - g_{\hat{k}-1}) \right] \\
&= \gamma_{\hat{k}} g_{\hat{k}+1}^T (g_{\hat{k}-1} - g_{\hat{k}-2}) = \gamma_{\hat{k}} [(1 - \beta_{\hat{k}}) g_{\hat{k}-1} + \beta_{\hat{k}} r_{\hat{k}}]^T (g_{\hat{k}-1} - g_{\hat{k}-2}), \quad (3.3)
\end{aligned}
$$

where $\gamma_{\hat{k}} = (\beta_{\hat{k}-1} - 1)/(\alpha_{\hat{k}-1} \beta_{\hat{k}-1})$ is a well-defined positive number. Finally, from (a) in Lemma 4 we have that $g_{\hat{k}-1}^T (g_{\hat{k}-1} - g_{\hat{k}-2}) = 0$ and also that $g_{\hat{k}}^T (g_{\hat{k}-1} - g_{\hat{k}-2}) = 0$, and hence using (3.3) combined with Step 6, Lemma 2, and the inductive hypothesis, yields

$$
\begin{aligned}
g_{\hat{k}+1}^T A g_{\hat{k}-1} &= \gamma_{\hat{k}} \beta_{\hat{k}} r_{\hat{k}}^T (g_{\hat{k}-1} - g_{\hat{k}-2}) = \gamma_{\hat{k}} \beta_{\hat{k}} (g_{\hat{k}} - \alpha_{\hat{k}} A g_{\hat{k}})^T (g_{\hat{k}-1} - g_{\hat{k}-2}) \\
&= \gamma_{\hat{k}} \beta_{\hat{k}} \alpha_{\hat{k}} g_{\hat{k}}^T A (g_{\hat{k}-2} - g_{\hat{k}-1}) = \gamma_{\hat{k}} \beta_{\hat{k}} \alpha_{\hat{k}} [g_{\hat{k}}^T A g_{\hat{k}-2} - g_{\hat{k}}^T A g_{\hat{k}-1}] = 0,
\end{aligned}
$$

and (i) is established for all $k \geq 2$ and for $-1 \leq j \leq k - 2$.

Concerning (ii), for $-1 \leq j \leq \hat{k} - 1$, using Step 9 of the DWGM algorithm, Lemma 2, (i) which has now been established, and that $\beta_k > 1$ for all k, we obtain

$$r_{\hat{k}+1}^T A g_j = \frac{1}{\beta_{\hat{k}+1}} [g_{\hat{k}+2} + (\beta_{\hat{k}+1} - 1) g_{\hat{k}}]^T A g_j = \frac{1}{\beta_{\hat{k}+1}} g_{\hat{k}+2}^T A g_j + \frac{\beta_{\hat{k}+1} - 1}{\beta_{\hat{k}+1}} g_{\hat{k}}^T A g_j = 0,$$

and (ii) is also established. □

Summing up, combining (b) from Lemma 1 with (ii) from Theorem 5, it follows that $r_k^T A g_j = 0$ for $j = k$ and for $j \leq k - 2$. In other words, for all k, r_k is A-orthogonal to

8

all previous gradient vectors except to g_{k-1}. Moreover, combining Lemma 2 with (i) from Theorem 5, it follows that for all k, g_k is A-orthogonal to all previous gradient vectors, i.e., for all $k \geq 1$

$$g_k^T A g_j = 0 \quad \text{for all} \quad j \leq k - 1. \tag{3.4}$$

We are now ready to show the finite termination of the DWGM algorithm.

Theorem 6. *For any initial guess $x_0 \in \mathbb{R}^n$, Algorithm DWGM generates the iterates x_k, $k \geq 1$, such that $x_n = A^{-1}b$.*

Proof. From (3.4) we have that the n vectors g_k, $0 \leq k \leq n - 1$, form an A-orthogonal set, and hence they form a linearly independent set of n vectors in $\mathbb{R}^n$. Therefore, the next vector $g_n \in \mathbb{R}^n$ must be zero to be able to keep the A-orthogonality with all the previous gradient vectors. Thus, $Ax_n = b$ and hence $x_n = A^{-1}b$. ☐
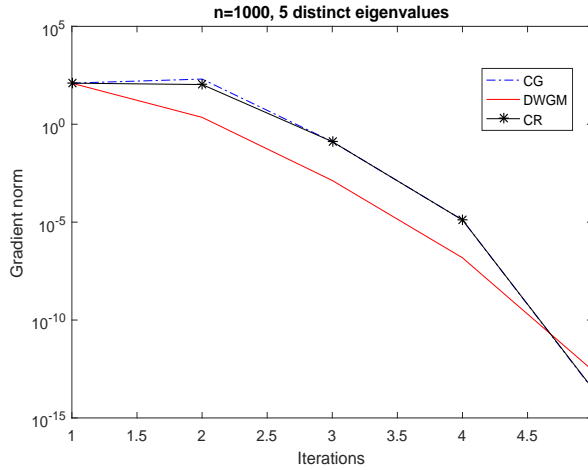

Figure 2: Convergence history of DWGM, CG, and CR, from the same initial point, for the minimization of a strictly convex quadratic function, when $n = 1000$ and the matrix A has only 5 distinct eigenvalues, equally distributed in the interval $[10, 1000]$, each one repeated 200 times.

It is worth noticing that in exact arithmetic the final termination of DWGM, as in the CG and CR methods, is related to the number of distinct eigenvalues of the matrix A, and not to the dimension of A. In Figure 2 this fact is illustrated on a strictly convex quadratic function, with $n = 1000$ and for which the matrix A has only 5 distinct eigenvalues. Indeed, we can observe that the three methods terminate in 5 iterations. A key observation, before establishing this important result, is that for each k the gradient vector g_k generated by DWGM belongs to the Krylov subspace $\mathcal{K}_{k+1}(A, g_0)$ (to be defined in our next lemma) that only depends on the matrix A and the initial gradient vector g_0.

Lemma 7. *In Algorithm DWGM, for all $k \geq 1$*

$$g_k \in \mathcal{K}_{k+1}(A, g_0) := span\{g_0, Ag_0, A^2 g_0, \ldots, A^k g_0\}.$$

Proof. For $k = 1$, using (a) in Lemma 1, we have $g_1 = r_0 = g_0 - \alpha_0 Ag_0$, and so $g_1 \in span\{g_0, Ag_0\}$. Let us now assume, by induction on k, that for all $1 \leq j \leq k$

$$g_j \in \mathcal{K}_{j+1}(A, g_0) = span\{g_0, Ag_0, A^2 g_0, \ldots, A^j g_0\},$$

and consider g_{k+1}. From Step 9 and Step 6 of the DWGM algorithm, we have

$$g_{k+1} = (1 - \beta_k)g_{k-1} + \beta_k r_k = (1 - \beta_k)g_{k-1} + \beta_k g_k - \beta_k \alpha_k Ag_k.$$

By the inductive hypothesis, $g_{k-1} \in \mathcal{K}_k(A, g_0)$, $g_k \in \mathcal{K}_{k+1}(A, g_0)$, and so $Ag_k \in \mathcal{K}_{k+2}(A, g_0)$. Consequently, $g_{k+1} \in \mathcal{K}_{k+2}(A, g_0)$ and the result is established. $\square$

Krylov spaces are closely related to polynomials. In fact, for any nonzero vector $z \in \mathbb{R}^n$ and any positive integer m, it is clear that

$$\mathcal{K}_m(A, z) := span\{z, Az, A^2 z, \ldots, A^{m-1} z\} = \{q(A)z \; : \; q \in \mathcal{P}_{m-1}\},$$

where $\mathcal{P}_{m-1}$ denotes the space of all polynomials of degree at most $m - 1$. Let us recall that the minimal polynomial of z with respect to A is the nonzero monic polynomial $\widehat{q}$ of lowest degree such that $\widehat{q}(A)z = 0$. The results stated in our next theorem are well-known and we present them here without a proof; for a complete discussion on the connection between Krylov spaces and polynomials see, e.g., [15, Ch. VI] and [16, Ch. 4].

Theorem 8. *The Krylov subspace $\mathcal{K}_m(A, z)$ is of dimension m if and only if the degree of the minimal polynomial $\widehat{q}$ of z with respect to A is greater than or equal to m. Moreover, if η is the degree of the the minimal polynomial $\widehat{q}$ of z with respect to A, then $\mathcal{K}_\eta(A, z)$ is invariant under A and $\mathcal{K}_m(A, z) = \mathcal{K}_\eta(A, z)$ for all $m \geq \eta$.*

We note that, based on Theorem 8, the degree of the minimal polynomial $\widehat{q}$ can also be characterized as the smallest positive integer η such that $\mathcal{K}_\eta(A, z) = \mathcal{K}_{\eta+1}(A, z)$. In particular, the relation between Krylov subspaces and the minimal polynomial of the initial gradient vector g_0 has played a fundamental role to study the finite termination of the CG and CR methods when the matrix A has only $p < n$ distinct eigenvalues; see e.g., [10, 14]. It will also play a key role to study the finite termination of the DWGM algorithm at iteration p.

Theorem 9. *If A has only $p < n$ distinct eigenvalues, then for any initial guess $x_0 \in \mathbb{R}^n$ Algorithm DWGM generates the iterates x_k, $k \geq 1$, such that $x_p = A^{-1}b$.*

Proof. Since A is symmetric and positive definite, the eigenvalues λ_i, $1 \leq i \leq p$, are positive and the associated eigenvectors v_i, $1 \leq i \leq n$, can be chosen to form an orthonormal set. Without loss of generality we can assume that the eigenvectors $\{v_1, \ldots, v_{i_1}\}$ are

10

associated with λ_1, the eigenvectors $\{v_{i_1+1}, \ldots, v_{i_2}\}$ with λ_2, and so on, until finally the eigenvectors $\{v_{i_{p-1}}, \ldots, v_n\}$ are associated with λ_p, where $1 \le i_1 < \cdots < i_{p-1} \le n$. Clearly, the set of eigenvectors form a basis in $\mathbb{R}^n$, and so there exist real scalars γ_i, $1 \le i \le n$, such that

$$g_0 = \sum_{j=1}^{i_1} \gamma_j v_j + \sum_{j=i_1+1}^{i_2} \gamma_j v_j + \cdots + \sum_{j=i_{p-1}}^{n} \gamma_j v_j = \sum_{i=1}^{p} \widehat{w}_i,$$

where $\widehat{w}_1 = \sum_{j=1}^{i_1} \gamma_j v_j$, $\widehat{w}_2 = \sum_{j=i_1+1}^{i_2} \gamma_j v_j$, and so on until $\widehat{w}_p = \sum_{j=i_{p-1}}^{n} \gamma_j v_j$. Hence, $g_0 \in span\{\widehat{w}_1, \widehat{w}_2, \ldots, \widehat{w}_p\}$. Moreover, for any $1 \le j \le p$, $A\widehat{w}_j = \lambda_j \widehat{w}_j$, and so the subspace $span\{\widehat{w}_1, \widehat{w}_2, \ldots, \widehat{w}_p\}$ is invariant under A. Furthermore, for any $0 \le k \le p-1$, $A^k g_0 = \sum_{i=1}^{p} \lambda_i^k \widehat{w}_i$ and we obtain the following column-wise matrix equality

$$\begin{bmatrix} | & | & & | \\ g_0 & Ag_0 & \cdots & A^{p-1}g_0 \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \widehat{w}_1 & \widehat{w}_2 & \cdots & \widehat{w}_p \\ | & | & & | \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{p-1} \\ \vdots & \vdots & & \vdots \\ 1 & \lambda_p & \cdots & \lambda_p^{p-1} \end{bmatrix}. \tag{3.5}$$

We note that the second matrix on the right hand side of (3.5) is a real $p \times p$ Vandermonde's matrix, whose determinant is given by $\prod_{1 \le i < j \le p}(\lambda_j - \lambda_i)$; see, e.g., [8, Sec. 6.1]. Since the p eigenvalues of A are distinct, we conclude that it is a nonsingular matrix. Hence, the column space of the two $n \times p$ matrices in (3.5) are equal, which implies that

$$\mathcal{K}_p(A, g_0) := span\{g_0, Ag_0, A^2 g_0, \ldots, A^{p-1}g_0\} = span\{\widehat{w}_1, \widehat{w}_2, \ldots, \widehat{w}_p\}.$$

Consequently, $\mathcal{K}_p(A, g_0)$ is also invariant under A, which in turn implies from Theorem 8 that the degree of the the the minimal polynomial $\widehat{q}$ of g_0 with respect to A is p. Therefore, combining (3.4) and Lemma 7, it follows that the only way for the vector g_p to be A-orthogonal to all the previous gradient vectors while being in $\mathcal{K}_p(A, g_0)$ is that $g_p = 0$. Thus, $Ax_p = b$ and we obtain that $x_p = A^{-1}b$. □

In addition to the A-orthogonality results shown in Theorem 5, we can also study the A-orthogonality of the current gradient g_k with all the previously explored search directions. As can be noticed from Step 8 of the DWGM algorithm, the search direction to move from x_k to the next iterate is not given explicitly, instead it is given the direction to move from x_{k-1} to x_{k+1}, which uses the auxiliary vector y_k. Nevertheless, we can consider the vector $x_k - x_{k-1}$ as the search direction to move from x_{k-1} to x_k. Notice that for any $1 \le j \le k$,

$$g_k^T A(x_j - x_{j-1}) = g_k^T[(Ax_j - b) - (Ax_{j-1} - b)] = g_k^T(g_j - g_{j-1}).$$

Using this equality, our next result establishes the mentioned A-orthogonality.

Theorem 10. *Algorithm DWGM generates the sequences $\{g_k\}$ such that for $k \ge 2$*

$$g_k^T(g_j - g_{j-1}) = 0 \quad \text{for all} \quad 1 \le j \le k. \tag{3.6}$$

11

Proof. Let us notice, from (e) in Lemma 3, that $g_2^T(g_1 - g_0) = 0$, and so the result is obtained for $k = 2$. Let us now assume, by induction on k, that (3.6) holds up to $k = \hat{k} \geq 3$, and consider the next iteration. Hence, we need to show that $g_{\hat{k}+1}^T(g_j - g_{j-1}) = 0$ for all $-1 \leq j \leq \hat{k} + 1$.

When $j = \hat{k}$, the result follows directly from (e) in Lemma 3, and when $j = \hat{k} + 1$, the result follows directly from (a) in Lemma 4. Now, if $j \leq \hat{k} - 2$, using Steps 6 and 9 of the DWGM algorithm, the inductive hypothesis on (3.6), Lemma 2, and (i) from Theorem 5, we have

$$
\begin{aligned}
g_{\hat{k}+1}^T(g_j - g_{j-1}) &= [(1 - \beta_{\hat{k}})g_{\hat{k}-1} + \beta_{\hat{k}}r_{\hat{k}}]^T(g_j - g_{j-1}) = \beta_{\hat{k}}r_{\hat{k}}^T(g_j - g_{j-1}) \\
&= \beta_{\hat{k}}(g_{\hat{k}} - \alpha_{\hat{k}}Ag_{\hat{k}})^T(g_j - g_{j-1}) = -\alpha_{\hat{k}}\beta_{\hat{k}}g_{\hat{k}}^T A(g_j - g_{j-1}) \\
&= -\alpha_{\hat{k}}\beta_{\hat{k}}[(g_{\hat{k}}^T Ag_j) - (g_{\hat{k}}^T Ag_{j-1})] = 0.
\end{aligned}
$$

Finally, when $j = \hat{k} - 1$, using Steps 6 (twice) and 9 of the DWGM algorithm, (d) from Lemma 1, Lemma 2, and (i) from Theorem 5, we obtain

$$
\begin{aligned}
g_{\hat{k}+1}^T(g_{\hat{k}-1} - g_{\hat{k}-2}) &= g_{\hat{k}+1}^T(g_{\hat{k}-1} - r_{\hat{k}} + r_{\hat{k}} - g_{\hat{k}-2}) = g_{\hat{k}+1}^T(r_{\hat{k}} - g_{\hat{k}-2}) \\
&= g_{\hat{k}+1}^T(g_{\hat{k}} - \alpha_{\hat{k}}Ag_{\hat{k}} - g_{\hat{k}-2}) = g_{\hat{k}+1}^T(g_{\hat{k}} - g_{\hat{k}-2}) \\
&= g_{\hat{k}+1}^T(g_{\hat{k}-2} + \beta_{\hat{k}-1}(r_{\hat{k}-1} - g_{\hat{k}-2}) - g_{\hat{k}-2}) = \beta_{\hat{k}-1}g_{\hat{k}+1}^T(r_{\hat{k}-1} - g_{\hat{k}-2}) \\
&= \beta_{\hat{k}-1}g_{\hat{k}+1}^T(g_{\hat{k}-1} - g_{\hat{k}-2} - \alpha_{\hat{k}-1}Ag_{\hat{k}-1}) = \beta_{\hat{k}-1}g_{\hat{k}+1}^T(g_{\hat{k}-1} - g_{\hat{k}-2}).
\end{aligned}
$$

Therefore, $(\beta_{\hat{k}-1} - 1)g_{\hat{k}+1}^T(g_{\hat{k}-1} - g_{\hat{k}-2}) = 0$. Since $\beta_k > 1$ for all $k \geq 1$ ((c) in Lemma 4), we conclude that $g_{\hat{k}+1}^T(g_{\hat{k}-1} - g_{\hat{k}-2}) = 0$, and (3.6) is established. □

Let us recall that the step length α_k is obtained in the DWGM algorithm to guarantee that the gradient norm is minimized along the negative gradient direction to obtain y_k; see [13, p. 731]. Our next result establishes, using Theorem 10, that the gradient norm at iteration k actually attains the minimum possible value on the linear variety (subspace if $x_0 = 0$) of dimension k generated by all the search directions that has been explored so far:

$$
V_k = \{x \in \mathbb{R}^n : x = x_0 + \sum_{j=1}^{k} \eta_j(x_j - x_{j-1}) \text{ and } \eta_j \in \mathbb{R}, \text{ for } 1 \leq j \leq k\}.
$$

In that sense, the gradient norm in the DWGM algorithm plays a similar role to the one played by the objective function in the CG method. Indeed, in the CG method (starting at $x_0 = 0$) the step length is chosen to guarantee that the function value $f(x)$ is minimized along the k-th search direction, but in reality based on the orthogonality of the gradient with all the previous search directions, the function value attains the minimum value on the entire explored subspace.

Corollary 11. *In Algorithm DWGM, for all $k \geq 1$, the iterate x_k is obtained such that $\|g_k\|_2$ is the minimum possible value of $\|\nabla f(x)\|_2$ on V_k.*

Proof. Let us notice that the minimization of $\|\nabla f(x)\|_2 = \|Ax - b\|_2$ subject to $x \in V_k$ is equivalent to the unconstrained minimization of

$$G(\eta) \equiv \|A(x_0 + \sum_{j=1}^{k} \eta_j(x_j - x_{j-1})) - b\|_2^2 = \|\sum_{j=1}^{k} \eta_j(g_j - g_{j-1}) + g_0\|_2^2, \qquad (3.7)$$

where $\eta = [\eta_1, \ldots, \eta_k]^T \in \mathbb{R}^k$. Notice that the objective function $G(\eta)$ is clearly a strictly convex function in $\mathbb{R}^k$. Hence, (3.7) has a unique solution, say $\eta^* \in \mathbb{R}^k$, and let us define $R(\eta^*) = (\sum_{j=1}^{k} \eta_j^*(g_j - g_{j-1}) + g_0) \in \mathbb{R}^n$. Since $G(\eta)$ is strictly convex, the necessary optimality conditions

$$\frac{\partial G(\eta)}{\partial \eta_j} = R(\eta^*)^T(g_j - g_{j-1}) = 0, \quad \text{for} \ \ 1 \leq j \leq k,$$

are also sufficient. Hence, it follows that $R(\eta^*)$ is orthogonal to the subspace generated by the vectors $[g_k - g_{k-1}, \ldots, g_1 - g_0]$. From Theorem 10 we have that g_k is also orthogonal to the subspace generated by $[g_k - g_{k-1}, \ldots, g_1 - g_0]$. Moreover, notice that choosing $\eta \in \mathbb{R}^k$ such that $\eta_j = 1$ for $1 \leq j \leq k$, we obtain that $x_k \in V_k$ and $R(\eta) = g_k$. Therefore, by the uniqueness of the solution of (3.7) we obtain that $g_k = R(\eta^*) \in \mathbb{R}^n$. Using now the equivalence of the two minimization problems stated above, we have that the iterate x_k in algorithm DWGM can be written as

$$x_k = x_0 + \sum_{j=1}^{k} \eta_j^*(x_j - x_{j-1}),$$

and the result is established. $\square$

4 Conclusions and perspectives

We have discussed and established several properties of the DWGM algorithm, originally developed and analyzed in [13], which add understanding to the surprisingly good behavior of the method. In particular, we have shown the A-orthogonality of the gradient vector at the current iteration with all the previous gradient vectors, which yields the finite termination of the method for the minimization of convex quadratics. We have also established the A-orthogonality of the gradient vector at the current iteration with all the previously explored directions, including the current one, which shows that the method guarantees at each iteration that the norm of the current gradient is optimal on the entire explored linear variety. We have also studied the finite termination in $p < n$ iterations when the $n \times n$ Hessian matrix has only p distinct eigenvalues, as it also happens for the

13

CG and CR methods. This result clearly motivates the use of preconditioning strategies when solving large-scale symmetric and positive definite linear systems.

An advantage of the DWGM algorithm is that it does not impose any of the established A-orthogonality results in its algorithmic steps, and also that it uses a step length α_k which does not require an exact line search, and as a consequence its extension to the local minimization of non-quadratic functions is appealing, as observed by Oviedo [13]. Another advantage for the possible extension of the DWGM algorithm to the non-quadratic case is the tendency to outperform the CG method when low accuracy in the gradient norm is required, which could be a key issue in practical applications.

References

[1] Brezinski C. [1996], *Variations on richardson's method and acceleration*, Bull. Belg.Math. Soc., 3(5): 33–44.

[2] Birgin E.G., Martínez J.M., and Raydan M. [2014], *Spectral projected gradient methods: review and perspectives*, J. Stat. Softw., 60(3): 1–21.

[3] Dai Y.-H., Huang Y., and Liu X.-W. [2019], *A family of spectral gradient methods for optimization*, Comput. Optim. Appl., 74: 43–65.

[4] De Asmundis R., Di Serafino D., Riccio F., and Toraldo G. [2013], *On spectral properties of steepest descent methods*, IMA J. Numer. Anal., 33(4): 1416–1435.

[5] De Asmundis R., Di Serafino D., Hager W.W., Toraldo G., and Zhang H. [2014], *An efficient gradient method using the Yuan steplength*, Comput. Optim. Appl., 59(3): 541–563.

[6] Di Serafino D., Ruggiero V., Toraldo G., and Zanni L. [2018], *On the steplength selection in gradient methods for unconstrained optimization*, Appl. Math. Comput., 318: 176–195.

[7] Friedlander A. , Martínez J.M., Molina B., and Raydan M. [1999], *Gradient method with retards and generalizations*, SIAM J. Numer. Anal., 36(1): 275–289.

[8] Horn R. A. and Johnson C. R. [1991], *Topics in Matrix Analysis*, Cambridge University Press.

[9] Huang Y., Dai Y.-H., Liu X.-W., and Zhang H. [2020], *Gradient methods exploiting spectral properties*, Optim. Meth. Soft., to appear, DOI: 10.1080/10556788.2020.1727476.

[10] Kelley T. [1995], *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia.

[11] Lamotte J.-L., Molina B., and Raydan M. [2002], *Smooth and adaptive gradient method with retards*, Math. Comput. Model., 36(9-10): 1161–1168.

[12] Luengo F. and Raydan M. [2003], *Gradient Method with Dynamical Retards for Large-Scale Optimization Problems*, Transactions on Numerical Analysis (ETNA), 16: 186–193.

[13] Oviedo-Leon H. F. [2019], *A delayed weighted gradient method for strictly convex quadratic minimization*, Comput. Optim. Appl., 74:729–746.

[14] Saad Y. [2010], *Iterative Methods for Sparse Linear Systems*, 2nd ed.; SIAM, Philadelphia.

[15] Saad Y. [2011], *Numerical Methods for Large Eigenvalue Problems*, 2nd ed.; SIAM, Philadelphia.

[16] Stewart G.W. [2001], *Matrix Algorithms, Volume II: Eigensystems*, SIAM, Philadelphia.