A unified convergence bound for conjugate gradient and accelerated gradient[*]

Sahar Karimi[†] Stephen Vavasis[‡]

May 1, 2016

Abstract

Nesterov's accelerated gradient method for minimizing a smooth strongly convex function f is known to reduce $f(\mathbf{x}_k) - f(\mathbf{x}^*)$ by a factor of $\epsilon \in (0,1)$ after $k \geq O(\sqrt{L/\ell}\log(1/\epsilon))$ iterations, where ℓ, L are the two parameters of smooth strong convexity. Furthermore, it is known that this is the best possible complexity in the function-gradient oracle model of computation. The method of linear conjugate gradients (CG) also satisfies the same complexity bound in the special case of strongly convex quadratic functions, but in this special case it is faster than the accelerated gradient method.

Despite similarities in the algorithms and their asymptotic convergence rates, the conventional analyses of the two methods are nearly disjoint. The purpose of this note is provide a single quantity that decreases on every step at the correct rate for both algorithms. Our unified bound is based on a potential similar to the potential in Nesterov's original analysis.

As a side benefit of this analysis, we provide a direct proof that conjugate gradient converges in $O(\sqrt{L/\ell}\log(1/\epsilon))$ iterations. In contrast, the traditional indirect proof first establishes this result for the Chebyshev algorithm, and then relies on optimality of conjugate gradient to show that its iterates are at least as good as Chebyshev iterates. To the best of our knowledge, ours is the first direct proof of the convergence rate of linear conjugate gradient in the literature.

1 Conjugate gradient

The method of conjugate gradients (CG) was introduced by Hestenes and Stiefel [7] for minimizing strongly convex quadratic functions of the form $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}/2 - \mathbf{b}^T\mathbf{x}$, where

[*] Supported in part by a grant from the U. S. Air Force Office of Scientific Research and in part by a Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

[†] Department of Combinatorics & Optimization, University of Waterloo, 200 University Ave. W., Waterloo, ON, N2L 3G1, Canada, sahar.karimi@gmail.com.

[‡] Department of Combinatorics & Optimization, University of Waterloo, 200 University Ave. W., Waterloo, ON, N2L 3G1, Canada, vavasis@uwaterloo.ca.

A is a symmetric positive definite matrix. We refer to this algorithm as "linear conjugate gradients."

There is a significant body of work on gradient methods for more general smooth, strongly convex functions. We say that a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$ is *smooth, strongly convex* [8] if there exist two scalars $L \geq \ell > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\ell \|\mathbf{x} - \mathbf{y}\|^2/2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \leq L\|\mathbf{x} - \mathbf{y}\|^2/2. \tag{1}$$

This is equivalent to assuming convexity and lower and upper Lipschitz constants on the gradient:

$$\ell\|\mathbf{x} - \mathbf{y}\| \leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

Nemirovsky and Yudin [12] proposed a method for minimizing smooth strongly convex functions requiring $k = O(\sqrt{L/l}\log(1/\epsilon))$ iterations to produce an iterate $\mathbf{x}_k$ such that $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon(f(\mathbf{x}_0) - f(\mathbf{x}^*))$, where $\mathbf{x}^*$ is the optimizer (necessarily unique under the assumptions made). A drawback of their method is that it requires an expensive two-dimensional optimization on each iteration. Nesterov [13] proposed another method, nowadays known as the "accelerated gradient method," which achieves the same optimal complexity that requires a single function and gradient evaluation on each iteration.

The accelerated gradient method, although optimal in theory, can be slow in practice. For example, in the case of quadratic function, computational testing shows that it is substantially slower than linear conjugate gradients. In the special case of strongly convex quadratic functions, the conjugate gradient has already been known to satisfy the same asymptotic bound since the 1960s.

Although the two methods satisfy the same asymptotic bound, the analyses of the two methods are completely different. In the case of accelerated gradient, there are two analyses by Nesterov [13, 14]. There is also a recent analysis of a variant of accelerated gradient [2], which views it as a kind of ellipsoid method. (This variant apparently requires exact line search.)

In the case of linear conjugate gradient, we are aware of no direct analysis of the algorithm. By "direct," we mean an analysis of $f(\mathbf{x}_k) - f(\mathbf{x}^*)$ using the recurrence inherent in CG. Instead, the standard analysis proves that another iterative method, for example Chebyshev iteration [5] or the heavy-ball iteration [17, 1] achieves reduction of $\left(1 - O(\sqrt{\ell/L})\right)$ per iteration. Then one appeals to the optimality of the CG iterate in the Krylov space generated by all of these methods to claim that the CG iterate must be at least as good as the others.

This paper is devoted to establishing a one-step convergence bound that applies to both methods. The one-step convergence bound has the form $\Psi_{k+1} \leq \Psi_k/(1 + \sqrt{\ell/L})$ for $k = 1, 2, \ldots$, where Ψ_k is a potential defined by (22). This potential involves both certain perturbed distance from the current iterate to the optimizer and the objective function residual. It should be noted that for the accelerated gradient method, neither the sequence $\|\mathbf{x}_k - \mathbf{x}^*\|$ nor $f(\mathbf{x}_k) - f(\mathbf{x}^*)$ is monotonically decreasing with k. Both of these sequences decrease monotonically for conjugate gradient (refer to (45) and (47) below), but neither decreases at the rate $1/(1 + \sqrt{\ell/L})$ on every step. Instead, the rate of decrease of these quantities (both in theory and in practice) is erratic. Thus, it is not obvious that there is a combination of these quantities that decreases at the proper rate on every iteration for both algorithms.

The $k = 0$ case of (22) is

$$(\ell/2)\Psi_0 = (\ell/2)\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + (f(\mathbf{x}_0) - f(\mathbf{x}^*)),$$

while $(\ell/2)\Psi_k \le f(\mathbf{x}_k) - f(\mathbf{x}^*)$. Furthermore, we show below that $\Psi_{k+1} \le \Psi_k/(1+\sqrt{\ell/L})$ for $k = 1, 2, \ldots$ and $\Psi_1 \le \Psi_0$. The consequence of all these bounds is the following theorem.

Theorem 1 *Let $f(\mathbf{x})$ be a strongly convex smooth function with convexity parameters ℓ, L. Then the accelerated gradient method produces a sequence of iterates $\mathbf{x}_k$ such that*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le C_0 \left(1 + \sqrt{\frac{\ell}{L}}\right)^{-(k-1)}, \tag{2}$$

where $\mathbf{x}^*$ is the (necessarily unique) optimizer and $C_0 = (\ell/2)\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + f(\mathbf{x}_0) - f(\mathbf{x}^*)$. When applied to a quadratic function, the conjugate gradient method produces a sequence satisfying this bound.

Note that $(1 + \sqrt{\ell/L})^{-1} \le (1 - \sqrt{\ell/(4L)})$, so (2) implies the usual theorem except for a constant factor. Note that in the $k = 0$ case, we establish only $\Psi_1 \le \Psi_0$ instead of the stronger $\Psi_{k+1} \le \Psi_k/(1 + \sqrt{\ell/L})$, which is valid for $k = 1, 2, \ldots$. This explains why the exponent in Theorem 1 is $k - 1$ rather than k.

In Section 2, we review the accelerated gradient method. In Section 3, we review conjugate gradient and convergence bound. In Section 4 we propose a single algorithmic framework that unifies both algorithms. Then, in the main technical sections of this article, Sections 5–7, we present our unified analysis of the two algorithms, which is an extension of the potential-function approach used in Nesterov's original analysis. Finally, in Section 8, we discuss some possible consequences and future directions made possible by the unified bound.

2 Accelerated gradient method

Following the treatment in his book [14] Nesterov's accelerated method can be described as follows. Given a strongly convex $f : \mathbb{R}^n \to \mathbb{R}$ with strong convexity parameters L, ℓ, one uses the recurrence:

```
Accelerated gradient method
$\mathbf{x}_0 :=$ arbitrary
for $k := 0, 1, 2, \ldots$
    $\mathbf{y}_{k+1} := \mathbf{x}_k + \theta_k \mathbf{s}_k$                    (3)
    $\mathbf{x}_{k+1} := \mathbf{y}_{k+1} - \nabla f(\mathbf{y}_{k+1})/L$       (4)
    $\mathbf{s}_{k+1} := \mathbf{x}_{k+1} - \mathbf{x}_k$                    (5)
end
```

In (3) when $k = 0$, $\mathbf{s}_0$ is undefined and hence we define $\mathbf{y}_1 := \mathbf{x}_0$, and thus $\theta_0 = 0$. For $k \geq 1$, several choices of θ_k are valid; our analysis uses

$$\theta_k = \frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}. \tag{6}$$

3 Conjugate gradient method

The conjugate gradient method for minimizing $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}/2 - \mathbf{b}^T\mathbf{x}$, where A is a symmetric positive definite matrix, is due to Hestenes and Stiefel [7] and is as follows.

```
LCG method
$\mathbf{x}_0 :=$ arbitrary
$\mathbf{r}_0 := \mathbf{b} - A\mathbf{x}_0$
for $k := 0, 1, 2, \ldots,$
    $\beta_{k+1} := \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}$          (7)
    $\mathbf{p}_{k+1} := \beta_{k+1}\mathbf{p}_k + \mathbf{r}_k$          (8)
    $\alpha_{k+1} := \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1}}$          (9)
    $\mathbf{x}_{k+1} := \mathbf{x}_k + \alpha_{k+1}\mathbf{p}_{k+1}$          (10)
    $\mathbf{r}_{k+1} := \mathbf{r}_k - \alpha_{k+1}A\mathbf{p}_{k+1}$          (11)
end
```

When $k = 0$, $\mathbf{r}_{k-1}$ is undefined. Hence we disregard (7) for specifying β_1 and instead take $\beta_1 = 0$, which implies $\mathbf{p}_1 = \mathbf{r}_0$ in (8). It is apparent from this recurrence that $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k = -\nabla f(\mathbf{x}_k)$ for all k. Here are two other well known relationships from [7]:

$$\mathbf{p}_k^T\mathbf{r}_k = 0, \qquad \text{(HS 5:3c)} \tag{12}$$

$$\frac{1}{\alpha_k} \in [\lambda_{\min}(A), \lambda_{\max}(A)]. \qquad \text{(HS 5:12)} \tag{13}$$

Several monographs explain the method in detail from different points of view including Golub and Van Loan [5], Trefethen and Bau [19], Greenbaum [6] and Liesen and Strakos [11].

The best-known theorem regarding the convergence rate of conjugate gradient is due to Daniel [3] (but see [11] for a more comprehensive perspective):

Theorem 2 *For the above iteration,*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq 4\left(\frac{1 - \sqrt{\ell/L}}{1 + \sqrt{\ell/L}}\right)^{2k} (f(\mathbf{x}_0) - f(\mathbf{x}^*)).$$

Daniel's proof and all others known to us use the following line of reasoning. First, Daniel uses a known result that the Chebyshev method satisfies the bound above. Then he relies on the fact that the Chebyshev iterate $\mathbf{x}_k^{\mathrm{Ch}}$ lies in the affine space $\mathbf{x}_0+\mathrm{span}\{\mathbf{b}, A\mathbf{b}, \ldots, A^{k-1}\mathbf{b}\}$. On the other hand, conjugate gradient is known to produce the vector $\mathbf{x}_k$ that is the optimizer of f over this affine space. Therefore, the conjugate gradient iteration produces at least the same amount of reduction in f. The analysis of the convergence rate of conjugate gradient developed below does not rely on the optimality with respect to the Krylov space.

Daniel's theorem is tight in the sense that for every choice of $0 < \ell < L$ and k, there a matrix A and starting vector $\mathbf{b}$ such that the bound in the theorem is achieved to within constant factors. This follows from a much more general result of Nesterov [14], which states that the bound in Daniel's theorem is the best possible bound for any algorithm that uses the function-gradient oracle model. Linear conjugate gradient applied to convex quadratic functions is a member of this class of algorithms. However, for particular choices of A, much better behavior may be observed from linear conjugate gradient.

4 Unified algorithm

In this section, we consider the following iterative framework, which has three sequences of scalar parameters, θ_k, ν_k and π_k for $k = 0, 1, \ldots$.

```
Unified framework
$\mathbf{x}_0 :=$ arbitrary
for $k := 0, 1, 2, \ldots$
    $\mathbf{y}_{k+1} := \mathbf{x}_k + \theta_k \mathbf{s}_k$                                  (14)
    $\mathbf{x}_{k+1} := \mathbf{x}_k + \nu_k \mathbf{s}_k - \pi_k \nabla f(\mathbf{y}_{k+1})$     (15)
    $\mathbf{s}_{k+1} := \mathbf{x}_{k+1} - \mathbf{x}_k$                                  (16)
end
```

When $k = 0$, we leave $\mathbf{s}_0$ undefined and take $\mathbf{y}_1 := \mathbf{x}_0$ in (14) and $\mathbf{x}_1 := \mathbf{x}_0 - \pi_0 \nabla f(\mathbf{y}_1)$ in (15). This in turn means that we start with $\nu_0 = \theta_0 = 0$.

It is straightforward to observe that the accelerated gradient method is a special case of the unified framework if we make the identification

$$\nu_k^{\mathrm{AG}} \equiv \theta_k^{\mathrm{AG}} \equiv \frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}}$$

for $k \geq 1$ and $\pi_k^{\mathrm{AG}} = 1/L$ for all $k \geq 0$.

The LCG method can be derived as a special case of the unified framework as follows. First, take $\theta_k^{\mathrm{CG}} \equiv 0$ so that $\mathbf{y}_{k+1} \equiv \mathbf{x}_k$ for all k. Comparing (16) and (10) we see that

$$\mathbf{s}_k = \alpha_k \mathbf{p}_k. \tag{17}$$

Substituting (8) into (10) yields

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_{k+1}(\beta_{k+1}\mathbf{p}_k + \mathbf{r}_k) \\ &= \mathbf{x}_k + \alpha_{k+1}\left(\frac{\beta_{k+1}}{\alpha_k}\mathbf{s}_k - \nabla f(\mathbf{x}_k)\right).\end{aligned}$$

We recover this recurrence if we take

$$\nu_k^{\mathrm{CG}} \equiv \frac{\alpha_{k+1}\beta_{k+1}}{\alpha_k}, \qquad k = 1, 2, \ldots, \tag{18}$$

$$\pi_k^{\mathrm{CG}} \equiv \alpha_{k+1}, \qquad k = 0, 1, 2 \ldots \tag{19}$$

in (15).

In both LCG and accelerated gradient, the parameters satisfy the following relationships, which we assume for the rest of this paper:

$$\nu_k \geq \theta_k \geq 0\,(k = 0, 1, \ldots); \quad \nu_k > 0\,(k = 1, 2, \ldots); \quad \pi_k > 0\,(k = 0, 1, \ldots). \tag{20}$$

5 A potential for both algorithms

In this section we propose the common potential for both algorithms that decreases on every iteration. The main result we establish is:

$$C\Psi_{k+1} \leq \Psi_k \tag{21}$$

where

$$C = 1 + \sqrt{\ell/L}$$

and Ψ_k is a potential at step k:

$$\Psi_k = \|\mathbf{w}_k\|^2 + \frac{2}{\ell}(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \tag{22}$$

Here,

$$\mathbf{w}_k = \mathbf{x}_k + \rho_k\mathbf{s}_k - \mathbf{x}^*,$$

where ρ_k, $k = 0, 1, \ldots$, is an additional sequences of scalars defined below (see (30) and (44)), ℓ is the lower strong-convexity parameter, L is the upper parameter, and $\mathbf{x}^*$ is the minimizer of f. In fact, in the case of accelerated gradient, a slightly stronger bound of

$$C = 1 + \frac{1}{\sqrt{L/\ell} - 1}$$

is established. In the case $k = 0$, we define $\mathbf{w}_0 = \mathbf{x}_0 - \mathbf{x}^*$ (hence $\rho_0 = 0$).

A potential involving these two terms was proposed in [13], and our analysis may therefore be regarded as a variant of Nesterov's technique. (In [13], only the second term of Ψ_k is updated by a scalar from one iteration to the next.)

6 Analysis of accelerated gradient

We start by rewriting $\mathbf{w}_{k+1}$ and $\mathbf{w}_k$ in terms of $\mathbf{y}_{k+1}$, $\mathbf{s}_k$ and $\mathbf{x}^*$:

$$\begin{aligned}\mathbf{w}_k &= \mathbf{x}_k + \rho_k\mathbf{s}_k - \mathbf{x}^* \\ &= \mathbf{y}_{k+1} + (\rho_k - \theta_k)\mathbf{s}_k - \mathbf{x}^*, \quad \text{(by (14))}\end{aligned}$$

and

$$\begin{aligned}\mathbf{w}_{k+1} &= \mathbf{x}_{k+1} + \rho_{k+1}\mathbf{s}_{k+1} - \mathbf{x}^* \\ &= (1+\rho_{k+1})\mathbf{x}_{k+1} - \rho_{k+1}\mathbf{x}_k - \mathbf{x}^* \quad \text{(by (16))} \\ &= (1+\rho_{k+1})(\mathbf{y}_{k+1} + (\nu_k - \theta_k)\mathbf{s}_k - \pi_k\nabla f(\mathbf{y}_{k+1})) - \rho_{k+1}\mathbf{x}_k - \mathbf{x}^* \quad \text{(by (14) and (15))} \\ &= \mathbf{y}_{k+1} + \rho_{k+1}(\mathbf{y}_{k+1} - \mathbf{x}_k) + (1+\rho_{k+1})((\nu_k - \theta_k)\mathbf{s}_k - \pi_k\nabla f(\mathbf{y}_{k+1})) - \mathbf{x}^* \\ &= \mathbf{y}_{k+1} + (\rho_{k+1}\theta_k + (1+\rho_{k+1})(\nu_k - \theta_k))\mathbf{s}_k - (1+\rho_{k+1})\pi_k\nabla f(\mathbf{y}_{k+1}) - \mathbf{x}^* \quad \text{(by (14))} \\ &= \mathbf{y}_{k+1} - \mathbf{x}^* + ((1+\rho_{k+1})\nu_k - \theta_k)\mathbf{s}_k - (1+\rho_{k+1})\pi_k\nabla f(\mathbf{y}_{k+1}).\end{aligned}$$

We now let $\xi = \sqrt{C}$, which implies that the first term of $C\Psi_{k+1} - \Psi_k$ is of the form:

$$\|\xi\mathbf{w}_{k+1}\|^2 - \|\mathbf{w}_k\|^2 = (\xi\mathbf{w}_{k+1} - \mathbf{w}_k)^T(\xi\mathbf{w}_{k+1} + \mathbf{w}_k). \tag{23}$$

We expand the two factors separately using the previously developed expressions for $\mathbf{w}_{k+1}$ and $\mathbf{w}_k$:

$$\begin{aligned}\xi\mathbf{w}_{k+1} - \mathbf{w}_k &= \xi(\mathbf{y}_{k+1} - \mathbf{x}^* + ((1+\rho_{k+1})\nu_k - \theta_k)\mathbf{s}_k - (1+\rho_{k+1})\pi_k\nabla f(\mathbf{y}_{k+1})) \\ &\quad - (\mathbf{y}_{k+1} + (\rho_k - \theta_k)\mathbf{s}_k - \mathbf{x}^*) \\ &\equiv \mathbf{t}_1 + \mathbf{t}_2 - \mathbf{t}_3\end{aligned}$$

where

$$\begin{aligned}\mathbf{t}_1 &= (\xi - 1)(\mathbf{y}_{k+1} - \mathbf{x}^*), \\ \mathbf{t}_2 &= (\xi((1+\rho_{k+1})\nu_k - \theta_k) - (\rho_k - \theta_k))\mathbf{s}_k, \\ \mathbf{t}_3 &= \xi(1+\rho_{k+1})\pi_k\nabla f(\mathbf{y}_{k+1}).\end{aligned}$$

Here, the vectors $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ also depend on iteration k, but we omit writing this dependence since k is fixed for this part of the analysis. Similarly,

$$\begin{aligned}\xi\mathbf{w}_{k+1} + \mathbf{w}_k &= \xi(\mathbf{y}_{k+1} - \mathbf{x}^* + ((1+\rho_{k+1})\nu_k - \theta_k)\mathbf{s}_k - (1+\rho_{k+1})\pi_k\nabla f(\mathbf{y}_{k+1})) \\ &\quad + (\mathbf{y}_{k+1} + (\rho_k - \theta_k)\mathbf{s}_k - \mathbf{x}^*) \\ &\equiv \mathbf{u}_1 + \mathbf{u}_2 - \mathbf{u}_3\end{aligned}$$

where

$$\begin{aligned}\mathbf{u}_1 &= (\xi + 1)(\mathbf{y}_{k+1} - \mathbf{x}^*), \\ \mathbf{u}_2 &= (\xi((1+\rho_{k+1})\nu_k - \theta_k) + (\rho_k - \theta_k))\mathbf{s}_k, \\ \mathbf{u}_3 &= \xi(1+\rho_{k+1})\pi_k\nabla f(\mathbf{y}_{k+1}).\end{aligned}$$

Thus, (23) is rewritten $(\mathbf{t}_1+\mathbf{t}_2-\mathbf{t}_3)^T(\mathbf{u}_1+\mathbf{u}_2-\mathbf{u}_3)$. This expansion contains nine terms. Writing these and gathering like terms (and noting the simple identity $(a-b)(c+d)+(a+b)(c-d)=2ac-2bd$) yields:

$$\begin{aligned}
\mathbf{t}_1^T\mathbf{u}_1 &= (\xi^2-1)\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2, && (24)\\
\mathbf{t}_1^T\mathbf{u}_2+\mathbf{t}_2^T\mathbf{u}_1 &= 2(\xi^2((1+\rho_{k+1})\nu_k-\theta_k)-(\rho_k-\theta_k))(\mathbf{y}_{k+1}-\mathbf{x}^*)^T\mathbf{s}_k, && (25)\\
\mathbf{t}_2^T\mathbf{u}_2 &= (\xi^2((1+\rho_{k+1})\nu_k-\theta_k)^2-(\rho_k-\theta_k)^2)\|\mathbf{s}_k\|^2, && (26)\\
-\mathbf{t}_1^T\mathbf{u}_3-\mathbf{t}_3^T\mathbf{u}_1 &= 2\xi^2(1+\rho_{k+1})\pi_k(\mathbf{x}^*-\mathbf{y}_{k+1})^T\nabla f(\mathbf{y}_{k+1}), && (27)\\
-\mathbf{t}_2^T\mathbf{u}_3-\mathbf{t}_3^T\mathbf{u}_2 &= -2\xi^2((1+\rho_{k+1})\nu_k-\theta_k)(1+\rho_{k+1})\pi_k\mathbf{s}_k^T\nabla f(\mathbf{y}_{k+1}), && (28)\\
\mathbf{t}_3^T\mathbf{u}_3 &= \xi^2(1+\rho_{k+1})^2\pi_k^2\|\nabla f(\mathbf{y}_{k+1})\|^2. && (29)
\end{aligned}$$

For accelerated gradients, we use a constant value for ρ_k (independent of k) that is analogous to the choice in [13], namely,

$$\rho_k=\sqrt{L/\ell}-1 \quad \text{for } k=1,2,\ldots. \tag{30}$$

Assume for now that $k\ge 1$; the $k=0$ case is considered separately below. The inner product $(\mathbf{y}_{k+1}-\mathbf{x}^*)^T\mathbf{s}_k$ in (25) appears difficult to bound in the case of accelerated gradient, so we define the scalar ξ^2 $(=C)$ to ensure that the term $\mathbf{t}_1^T\mathbf{u}_2+\mathbf{t}_2^T\mathbf{u}_1$ is zero, namely,

$$\begin{aligned}
\xi^2 &= \frac{\rho_k-\theta_k}{(1+\rho_{k+1})\nu_k-\theta_k}\\
&= \frac{\sqrt{L/\ell}-1-(\sqrt{L/\ell}-1)/(\sqrt{L/\ell}+1)}{(\sqrt{L/\ell}-1)\cdot(\sqrt{L/\ell}-1)/(\sqrt{L/\ell}+1)}\\
&= 1+\frac{1}{\sqrt{L/\ell}-1}. && (31)
\end{aligned}$$

Note that this implies $C\ge 1+\sqrt{\ell/L}$, so that (2) will be established for this choice of ξ.

Next, we rewrite the remaining terms of (24)–(29) based on these choices for the scalars:

$$\begin{aligned}
\mathbf{t}_1^T\mathbf{u}_1 &= \frac{1}{\sqrt{L/\ell}-1}\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2, && (32)\\
\mathbf{t}_2^T\mathbf{u}_2 &= -\frac{\sqrt{L/\ell}(\sqrt{L/\ell}-1)^2}{(\sqrt{L/\ell}+1)^2}\|\mathbf{s}_k\|^2, && (33)\\
-\mathbf{t}_1^T\mathbf{u}_3-\mathbf{t}_3^T\mathbf{u}_1 &= \frac{2}{(\sqrt{L/\ell}-1)\ell}(\mathbf{x}^*-\mathbf{y}_{k+1})^T\nabla f(\mathbf{y}_{k+1}), && (34)\\
-\mathbf{t}_2^T\mathbf{u}_3-\mathbf{t}_3^T\mathbf{u}_2 &= -\frac{2}{\ell}\cdot\frac{\sqrt{L/\ell}-1}{\sqrt{L/\ell}+1}\cdot\mathbf{s}_k^T\nabla f(\mathbf{y}_{k+1}), && (35)\\
\mathbf{t}_3^T\mathbf{u}_3 &= \frac{1}{L^{1/2}\ell^{3/2}(\sqrt{L/\ell}-1)}\|\nabla f(\mathbf{y}_{k+1})\|^2. && (36)
\end{aligned}$$

We analyze the sum of (34), (35), and (36) together:

$$-\mathbf{t}_1^T\mathbf{u}_3 - \mathbf{t}_3^T\mathbf{u}_1 - \mathbf{t}_2^T\mathbf{u}_3 - \mathbf{t}_3^T\mathbf{u}_2 + \mathbf{t}_3^T\mathbf{u}_3 = \frac{2}{(\sqrt{L/\ell}-1)\ell}\cdot t_4 \tag{37}$$

where

$$t_4 = (\mathbf{x}^* - \mathbf{y}_{k+1})^T\nabla f(\mathbf{y}_{k+1}) - \frac{(\sqrt{L/\ell}-1)^2}{\sqrt{L/\ell}+1}\mathbf{s}_k^T\nabla f(\mathbf{y}_{k+1}) + \frac{1}{2\sqrt{L\ell}}\|\nabla f(\mathbf{y}_{k+1})\|^2. \tag{38}$$

To analyze t_4 requires two more bounds. First, by (1), for any $\mathbf{x}\in\mathbb{R}^n$,

$$f(\mathbf{x}) \geq f(\mathbf{y}_{k+1}) + \nabla f(\mathbf{y}_{k+1})^T(\mathbf{x}-\mathbf{y}_{k+1}) + \frac{\ell}{2}\|\mathbf{x}-\mathbf{y}_{k+1}\|^2. \tag{39}$$

We also need a bound on the descent made per step. We use the well known bound

$$f\left(\mathbf{y}_{k+1}\right) - f(\mathbf{y}_{k+1} - \nabla f(\mathbf{y}_{k+1})/L) \geq \|\nabla f(\mathbf{y}_{k+1})\|^2/(2L). \tag{40}$$

This follows by writing the left-hand side $f(\mathbf{y}_{k+1}) - f(\mathbf{y}_{k+1}-\mathbf{d})$ as the line integral $\int_0^1 \nabla f(\mathbf{y}_{k+1}-t\mathbf{d})^T\mathbf{d}\,dt$ for the particular choice $\mathbf{d}=\nabla f(\mathbf{y}_{k+1})/L$, pulling out an additive term of $\|\nabla f(\mathbf{y}_{k+1})\|^2/L$ from the integrand, and then applying the Lipschitz condition.

Then the claimed bound is:

$$t_4 \leq (\sqrt{L/\ell}-1)(f(\mathbf{x}_k)-f(\mathbf{x}^*)) - \sqrt{L/\ell}(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)) - \frac{\ell}{2}\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2. \tag{41}$$

The following chain of inequalities starting from (38) establishes (41):

$$\begin{aligned}
t_4 &= (\mathbf{x}^*-\mathbf{y}_{k+1})^T\nabla f(\mathbf{y}_{k+1}) + (\sqrt{L/\ell}-1)(\mathbf{x}_k-\mathbf{y}_{k+1})^T\nabla f(\mathbf{y}_{k+1}) + \frac{\|\nabla f(\mathbf{y}_{k+1})\|^2}{2\sqrt{L\ell}} \\
&\qquad \text{(by (14))} \\
&\leq f(\mathbf{x}^*) - f(\mathbf{y}_{k+1}) - \frac{\ell}{2}\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2 + (\sqrt{L/\ell}-1)(f(\mathbf{x}_k)-f(\mathbf{y}_{k+1}) - \frac{\ell}{2}\|\mathbf{y}_{k+1}-\mathbf{x}_k\|^2) \\
&\quad + \frac{\|\nabla f(\mathbf{y}_{k+1})\|^2}{2\sqrt{L\ell}} \quad \text{(by (39))} \\
&\leq f(\mathbf{x}^*) - f(\mathbf{y}_{k+1}) - \frac{\ell}{2}\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2 + (\sqrt{L/\ell}-1)(f(\mathbf{x}_k)-f(\mathbf{y}_{k+1})) + \frac{\|\nabla f(\mathbf{y}_{k+1})\|^2}{2\sqrt{L\ell}} \\
&= f(\mathbf{x}^*) - \sqrt{L/\ell}\cdot f(\mathbf{y}_{k+1}) - \frac{\ell}{2}\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2 + (\sqrt{L/\ell}-1)f(\mathbf{x}_k) + \frac{\|\nabla f(\mathbf{y}_{k+1})\|^2}{2\sqrt{L\ell}} \\
&\leq f(\mathbf{x}^*) - \sqrt{L/\ell}\cdot\left(f\left(\mathbf{y}_{k+1}-\nabla f(\mathbf{y}_{k+1})/L\right) + \|\nabla f(\mathbf{y}_{k+1})\|^2/(2L)\right) \\
&\quad - \frac{\ell}{2}\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2 + (\sqrt{L/\ell}-1)f(\mathbf{x}_k) + \frac{\|\nabla f(\mathbf{y}_{k+1})\|^2}{2\sqrt{L\ell}} \quad \text{(by (40))} \\
&= f(\mathbf{x}^*) - \sqrt{L/\ell}\cdot f\left(\mathbf{y}_{k+1}-\nabla f(\mathbf{y}_{k+1})/L\right) - \frac{\ell}{2}\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2 + (\sqrt{L/\ell}-1)f(\mathbf{x}_k) \\
&= f(\mathbf{x}^*) - \sqrt{L/\ell}\cdot f(\mathbf{x}_{k+1}) - \frac{\ell}{2}\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2 + (\sqrt{L/\ell}-1)f(\mathbf{x}_k) \quad \text{(by (15))} \\
&= (\sqrt{L/\ell}-1)(f(\mathbf{x}_k)-f(\mathbf{x}^*)) - \sqrt{L/\ell}\cdot(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)) - \frac{\ell}{2}\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2
\end{aligned}$$

We now can finally analyze the bound on the first term of $C\Psi_{k+1}-\Psi_k$. We have an explicit formula for (24), we have forced (25) to be 0 by choice of ξ, and (33) is nonpositive. The remaining terms are captured in (37) and (41), so therefore

$$\begin{aligned}\|\xi\mathbf{w}_{k+1}\|^2-\|\mathbf{w}_k\|^2 \le{}& \frac{2}{(\sqrt{L/\ell}-1)\ell}\\ &\cdot\left[(\sqrt{L/\ell}-1)(f(\mathbf{x}_k)-f(\mathbf{x}^*))-\sqrt{L/\ell}\cdot(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*))\right]\\ &+\left[\frac{1}{\sqrt{L/\ell}-1}-\frac{2}{(\sqrt{L/\ell}-1)\ell}\cdot\frac{\ell}{2}\right]\|\mathbf{y}_{k+1}-\mathbf{x}^*\|^2. \qquad (42)\end{aligned}$$

Observe that the square-bracketed coefficient at the end of (42) is 0. Rearranging,

$$\begin{aligned}\|\xi\mathbf{w}_{k+1}\|^2+\frac{2}{(\sqrt{L/\ell}-1)\ell}\cdot\sqrt{L/\ell}\cdot(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*)) \le{}& \|\mathbf{w}_k\|^2\\ &+\frac{2}{(\sqrt{L/\ell}-1)\ell}\\ &\cdot(\sqrt{L/\ell}-1)(f(\mathbf{x}_k)-f(\mathbf{x}^*)),\end{aligned}$$

i.e.,

$$\xi^2\left[\|\mathbf{w}_{k+1}\|^2+\frac{2}{\ell}\cdot(f(\mathbf{x}_{k+1})-f(\mathbf{x}^*))\right]\le\|\mathbf{w}_k\|^2+\frac{2}{\ell}\cdot(f(\mathbf{x}_k)-f(\mathbf{x}^*)).$$

Thus, we have established (21) in the case $C=\xi^2$, which by (31) implies

$$C=1+\frac{1}{\sqrt{L/\ell}-1}. \qquad (43)$$

The case $k=0$ needs separate attention. First, by taking $\xi=1$, $\rho_0=0$, $\rho_1=\sqrt{L/\ell}-1$, $\theta_0=\nu_0=0$, we observe that (24), (25), (26) and (28) all vanish. Thus,

$$\begin{aligned}\|\mathbf{w}_1\|^2&=\|\mathbf{w}_0\|^2+\frac{2(\mathbf{x}^*-\mathbf{y}_1)^T\nabla f(\mathbf{y}_1)}{\sqrt{L\ell}}+\frac{\|\nabla f(\mathbf{y}_1)\|^2}{L\ell}\\ &\le\|\mathbf{w}_0\|^2+\frac{\|\nabla f(\mathbf{y}_1)\|^2}{L\ell}\end{aligned}$$

since the dropped term in the last line is nonpositive by convexity.

On the other hand,

$$\begin{aligned}\frac{2(f(\mathbf{x}_1)-f(\mathbf{x}^*))}{\ell}&=\frac{2(f(\mathbf{y}_1-\nabla f(\mathbf{y}_1)/L)-f(\mathbf{x}^*))}{\ell}\\ &\le\frac{2(f(\mathbf{y}_1)-\|\nabla f(\mathbf{y}_1)\|^2/(2L)-f(\mathbf{x}^*))}{\ell} \qquad \text{(by (40))}\\ &=\frac{2(f(\mathbf{x}_0)-\|\nabla f(\mathbf{y}_1)\|^2/(2L)-f(\mathbf{x}^*))}{\ell}\\ &=\frac{2(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{\ell}-\frac{\|\nabla f(\mathbf{y}_1)\|^2}{L\ell}.\end{aligned}$$

Adding the two preceding inequalities shows that $\Psi_1\le\Psi_0$. Thus, in this particular case, (21) does not necessarily hold for any $C>1$.

7 Convergence of conjugate gradient

We introduce the notation $F_k = 2(f(\mathbf{x}_k) - f(\mathbf{x}^*))$. We define

$$\rho_k = \frac{F_k}{\alpha_k \|\mathbf{r}_{k-1}\|^2}, \tag{44}$$

for $k = 1, 2, \ldots$ and $\rho_0 =$. For $k = 1, 2, \ldots$, this ρ_k has the special property that it is the optimizer of the optimization problem $\min\{\|\mathbf{x}_k + \rho\mathbf{s}_k - \mathbf{x}^*\| : \rho \in \mathbb{R}\}$, a property proved by [7] (see (6:8)). This property is not directly used in the upcoming analysis.

We also require the following result:

$$F_k - F_{k+1} = \alpha_{k+1}\|\mathbf{r}_k\|^2, \tag{45}$$

which follows from (9), (10), (12) and the fact that

$$f(\mathbf{x}_k + \mathbf{d}) = f(\mathbf{x}_k) - \mathbf{r}_k^T\mathbf{d} + \mathbf{d}^T A\mathbf{d}/2 \tag{46}$$

for any $\mathbf{d}$. It is also proven in [7, (6:1)].

We use two other equations from [7], the first of which is (6:5):

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 = \frac{(F_k + F_{k+1})\|\mathbf{p}_{k+1}\|^2}{\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1}}, \tag{47}$$

for $k = 0, 1, 2 \ldots$. Let us assume now that $k \geq 1$; the $k = 0$ case is considered below. The next equation, which holds for $k = 1, 2, \ldots$, is an unnumbered equation of [7, p. 417, col. 2]:

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_k + \rho_k\mathbf{s}_k - \mathbf{x}^*\|^2 = \frac{F_k^2\|\mathbf{p}_k\|^2}{\|\mathbf{r}_{k-1}\|^4}. \tag{48}$$

If we subtract (47) and the $k+1$ case of (48) from the k case of (48), and recalling the notation $\mathbf{w}_k = \mathbf{x}_k + \rho_k\mathbf{s}_k - \mathbf{x}^*$, we obtain

$$\|\mathbf{w}_{k+1}\|^2 - \|\mathbf{w}_k\|^2 = z_1 + z_2 + z_3 + z_4 \tag{49}$$

where

$$\begin{aligned}
z_1 &= \frac{F_k^2\|\mathbf{p}_k\|^2}{\|\mathbf{r}_{k-1}\|^4}, \\
z_2 &= -\frac{F_{k+1}^2\|\mathbf{p}_{k+1}\|^2}{\|\mathbf{r}_k\|^4}, \\
z_3 &= -\frac{F_k\|\mathbf{p}_{k+1}\|^2}{\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1}}, \\
z_4 &= -\frac{F_{k+1}\|\mathbf{p}_{k+1}\|^2}{\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1}}.
\end{aligned}$$

In order to simplify this sum, we make the following substitutions:

$$\begin{aligned} F_{k+1} &:= F_k - \alpha_{k+1}\|\mathbf{r}_k\|^2 && \text{(by (45))}, \\ \|\mathbf{r}_{k-1}\|^2 &:= \|\mathbf{r}_k\|^2/\beta_{k+1} && \text{(by (7))}, \\ \|\mathbf{p}_k\|^2 &:= (\|\mathbf{p}_{k+1}\|^2 - \|\mathbf{r}_k\|^2)/\beta_{k+1}^2 && \text{(by (8) and (12))} \end{aligned}$$

to obtain:

$$\begin{aligned} z_1 &= \frac{F_k^2(\|\mathbf{p}_{k+1}\|^2 - \|\mathbf{r}_k\|^2)}{\|\mathbf{r}_k\|^4}, \\ z_2 &= -\frac{(F_k^2 - 2F_k\alpha_{k+1}\|\mathbf{r}_k\|^2 + \alpha_{k+1}^2\|\mathbf{r}_k\|^4)\|\mathbf{p}_{k+1}\|^2}{\|\mathbf{r}_k\|^4}, \\ z_3 &= -\frac{F_k\|\mathbf{p}_{k+1}\|^2}{\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1}}, \\ z_4 &= -\frac{(F_k - \alpha_{k+1}\|\mathbf{r}_k\|^2)\|\mathbf{p}_{k+1}\|^2}{\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1}}. \end{aligned}$$

Now let us combine these terms, noting that the first term in z_1 cancels the first in z_2, and substituting $\alpha_{k+1} := \|\mathbf{r}_k\|^2/(\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1})$ (by (9)) to obtain

$$\begin{aligned} \|\mathbf{w}_{k+1}\|^2 - \|\mathbf{w}_k\|^2 &= -\frac{F_k^2}{\|\mathbf{r}_k\|^2} && \text{(term from } z_1) \\ &\quad + \frac{2F_k\|\mathbf{p}_{k+1}\|^2}{\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1}} - \frac{\|\mathbf{p}_{k+1}\|^2 \cdot \|\mathbf{r}_k\|^4}{(\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1})^2} && \text{(terms from } z_2) \\ &\quad - \frac{F_k\|\mathbf{p}_{k+1}\|^2}{\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1}} && \text{(term from } z_3) \\ &\quad - \frac{F_k\|\mathbf{p}_{k+1}\|^2}{\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1}} + \frac{\|\mathbf{r}_k\|^4 \cdot \|\mathbf{p}_{k+1}\|^2}{(\mathbf{p}_{k+1}^T A\mathbf{p}_{k+1})^2} && \text{(terms from } z_4) \\ &= -\frac{F_k^2}{\|\mathbf{r}_k\|^2}. \end{aligned} \tag{50}$$

It is also possible to obtain (50) from (24)–(29) with the choice $\xi = 1$.

Another helpful inequality is

$$\begin{aligned} \|\mathbf{w}_k\|^2 &\le \|\mathbf{x}_k - \mathbf{x}^*\|^2 && \text{(by (48))} \\ &\le (\mathbf{x}_k - \mathbf{x}^*)^T(A/\ell)(\mathbf{x}_k - \mathbf{x}^*) && \text{(since } \lambda_{\min}(A/\ell) \ge 1) \\ &= F_k/\ell. \end{aligned} \tag{51}$$

Now we establish (21):

$$\begin{aligned}
C\Psi_{k+1} - \Psi_k &= C(\|\mathbf{w}_{k+1}\|^2 + F_{k+1}/\ell) - (\|\mathbf{w}_k\|^2 + F_k/\ell) \\
&= (C-1)(\|\mathbf{w}_{k+1}\|^2 + F_{k+1}/\ell) + \|\mathbf{w}_{k+1}\|^2 - \|\mathbf{w}_k\|^2 \\
&\qquad + (F_{k+1} - F_k)/\ell \\
&= (C-1)(\|\mathbf{w}_{k+1}\|^2 + F_{k+1}/\ell) - F_k^2/\|\mathbf{r}_k\|^2 \\
&\qquad - \alpha_{k+1}\|\mathbf{r}_k\|^2/\ell && \text{(by (45) and (50))} \\
&\le (C-1)(\|\mathbf{w}_{k+1}\|^2 + F_{k+1}/\ell) - F_k^2/\|\mathbf{r}_k\|^2 \\
&\qquad - \|\mathbf{r}_k\|^2/(L\ell) && \text{(by (13))} \\
&\le (C-1)(\|\mathbf{w}_{k+1}\|^2 + F_{k+1}/\ell) - 2F_k/\sqrt{L\ell} && \text{(since } x^2+y^2 \ge 2xy\text{)} \\
&\le 2(C-1)F_{k+1}/\ell - 2F_k/\sqrt{L\ell} && \text{(by (51))} \\
&\le 2(C-1)F_k/\ell - 2F_k/\sqrt{L\ell} \\
&\le 0
\end{aligned}$$

provided that $2(C-1)/\ell \le 2/\sqrt{L\ell}$, i.e.,

$$C \le 1 + \sqrt{\frac{\ell}{L}}. \tag{52}$$

Thus, we take C equal to the right-hand side of the preceding inequality to establish (21).

Again, the $k = 0$ case needs special attention. For this case, as with accelerated gradient, we settle for the weaker inequality that $\Psi_1 \le \Psi_0$. This inequality holds for each of the two terms separately:

$$\begin{aligned}
\|\mathbf{w}_1\|^2 &\le \|\mathbf{x}_1 - \mathbf{x}^*\|^2 && \text{(by (48))} \\
&\le \|\mathbf{x}_0 - \mathbf{x}^*\|^2 && \text{(by (47))} \\
&= \|\mathbf{w}_0\|^2.
\end{aligned}$$

Also, $F_1 \le F_0$ by (45).

8 Discussion

The main point of this work is to prove Theorem 1 using the same convergence bound for accelerated gradient and conjugate gradient. The result in this paper was originally motivated by our consideration of nonlinear conjugate gradient.

The traditional extensions of linear CG to nonlinear CG for the general case of unconstrained optimization, e.g., the algorithms of Fletcher and Reeves [4] and Polak and Ribière [16] (see Nocedal and Wright [15] for an overview of these algorithms) are not optimal for minimizing strongly convex functions. Unlike accelerated gradient, there is no global complexity bound known for any nonlinear CG method even in the case of strongly convex functions. Indeed, Nemirovsky and Yudin argue that traditional nonlinear CG can perform even worse than steepest descent.

A partial unification the analyses of linear CG and accelerated gradient such as ours could point the way to development of a new nonlinear CG method. The new method would have two desirable properties: (1) it reduces to linear CG in the case of a quadratic function, and (2) it maintains the global convergence bound of accelerated gradient. Furthermore, such an algorithm would ideally be able to adapt between steps of the two algorithms even within the same problem. A preliminary proposal for a nonlinear CG like this was made in the PhD thesis of the first author [9], and will be the subject of ongoing work.

A second practical use of the unified analysis is the consideration of algorithms for minimizing a quadratic function using a modification of linear CG, such that the modification changes it into a nonlinear iteration. For example, several authors [18, 20] have considered the use of conjugate gradient methods in the case of noisy matrix-vector multiplication. Other authors, e.g., [10] have considered the possibility of changing the preconditioner from one iterate to the next. In both of these cases, optimality with respect to the Krylov space is no longer assured. However, it is possible that the bound in Theorem 1 may still hold. Since the analysis establishes Theorem 1 without relying on Krylov optimality, it may enable new analyses of such 'perturbed' conjugate gradient methods. This matter is also left for future work.

Another application of the LCG bound developed herein is to computational scientists developing new linear conjugate gradient methods (e.g., new preconditioners or new ways to compute matrix-vector products). Our bound not only directly shows the convergence rate claimed in Theorem 1, but more strongly it shows that the potential decreases by at least a fixed constant factor on each iteration. In a test run of any proposed new algorithm, it is possible to measure the potential developed herein and monitor its steady decrease. Any failure to exhibit the prescribed decrease would be an unambiguous indication that the method is failing due to some source of inexactness (e.g., roundoff error). In contrast, better known measures of LCG convergence can stagnate for many consecutive iterations, making it difficult to detect the impact of inexactness. We remark that in order to use our potential in this manner, it is of course necessary to know ℓ, L and the exact solution to the linear system at the outset, which is often the case in testing a new algorithm but obviously not in its practical use.

As for the theoretical content of this paper, it would be useful to simplify our analysis, which appears to complicated, and also to further unify the treatments of the two algorithms. Another useful development would be a potential that involves the term $\|\nabla f(\mathbf{x}_k)\|^2$. This is because, in practice, an algorithm does not have access to either $f(\mathbf{x}_k) - f(\mathbf{x}^*)$ or $\|\mathbf{x}_k - \mathbf{x}^*\|$, so the potential proposed here could not be evaluated by an algorithm to measure progress.

References

[1] D. P. Bertsekas. *Nonlinear programming (2nd edition)*. Athena Scientific, 1999.

[2] S. Bubeck, Y. T. Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. `http://arxiv.org/abs/1506.08187`, 2015.

[3] James W Daniel. The conjugate gradient method for linear and nonlinear operator equations. *SIAM Journal on Numerical Analysis*, 4(1):10–26, 1967.

[4] R. Fletcher and C. Reeves. Function minimization by conjugate gradients. *Computer Journal*, 7:149–154, 1964.

[5] G. H. Golub and C. F. Van Loan. *Matrix Computations, 2nd Edition.* Johns Hopkins University Press, Baltimore, 1989.

[6] A. Greenbaum. *Iterative methods for solving linear systems.* SIAM Publications, 1997.

[7] Magnus Rudolph Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.

[8] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis.* Springer, 2012.

[9] Sahar Karimi. *On the relationship between conjugate gradient and optimal first-order methods for convex optimization.* PhD thesis, University of Waterloo, 2014.

[10] A. V. Knyazev and I. Lashuk. Steepest descent and conjugate gradient methods with variable preconditioning. *SIAM J. Matrix Anal. Appl.*, 29:1267–1280, 2007.

[11] J. Liesen and Z. Strakos. *Krylov subspace methods: principles and analysis.* Oxford Science Publications, 2013.

[12] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* John Wiley and Sons, Chichester, 1983. Translated by E. R. Dawson from *Slozhnost' Zadach i Effektivnost' Metodov Optimizatsii*, 1979, Glavnaya redaktsiya fiziko-matematicheskoi literatury, Izdatelstva "Nauka".

[13] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR (translated as Soviet Math. Dokl.)*, 269(3):543–547, 1983.

[14] Y. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer, 2003.

[15] J. Nocedal and S. Wright. *Numerical Optimization, 2nd Edition.* Springer, New York, 2006.

[16] E. Polak and G. Ribière. Note sur la convergence de méthods de directions conjugées. *Revue Français d'informatique et de Recherche Opérationelle*, 3:35–43, 1969.

[17] R. Polyak. Modified barrier functions (theory and methods). *Mathematical Programming*, 54:177–222, 1992.

[18] V. Simoncini and D. Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. *SIAM J. Sci. Comput.*, 25:454–477, 2003.

[19] L. N. Trefethen and D. Bau. *Numerical Linear Algebra.* SIAM, 1997.

[20] J. van den Eshof and G. Sleijpen. Inexact Krylov subspace methods for linear systems. *SIAM J. Matrix Anal. Appl.*, 26:125–153, 2004.