Algorithms with Convergence to Second-Order Optimal Points

Daniel P. Robinson
Department of Applied Mathematics and Statistics
Johns Hopkins University

November 19, 2020
Outline

1. Introduction

2. Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3. Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4. Trust-Region Methods

5. Conclusions and Final Thoughts
Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, consider the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

(1)

- the case when $f$ is continuously differentiable (first-order methods)
- the case when $f$ is twice continuously differentiable (second-order methods)

Notation:

- $f(x_k) := f(x_k)$, $g_k := \nabla f(x_k)$, $H_k := \nabla^2 f(x_k)$
- We let $(\lambda_k, v_k)$ denote a left-most eigenpair of $H_k$ with $||v_k||_2 = 1$. 
Previous methods

- **line search algorithms**
  - \( x_{k+1} = x_k + \alpha_k p_k \)
  - \( p_k \) was a descent direction, i.e., \( g_k^T p_k < 0 \)
  - \( \alpha_k > 0 \) was chosen using some line search procedure, e.g., backtracking Armijo
  - if \( f \) is bounded below on the initial level set and finite termination does not happen, then
    \[
    \lim_{k \to \infty} \|g_k\| = 0
    \]
    thus, all limit points of \( \{x_k\} \) satisfy the **first-order** necessary optimality condition.
  - notice that if \( g_k = 0 \), then no \( p_k \) exists such that \( g_k^T p_k < 0 \).
  - thus, can’t prove that limit points of \( \{x_k\} \) for previous line search algorithms satisfy
    **second-order** optimality necessary conditions.

- **trust region algorithms**
  - \( x_{k+1} = x_k + s_k \)
  - \( s_k \) was the solution to a trust region subproblem
  - an appropriate step \( s_k \) found by adjusting trust-region radius
  - if \( f \) is bounded below on the initial level set and finite termination does not happen, then
    \[
    \lim_{k \to \infty} \|g_k\| = 0
    \]
    thus, all limit points of \( \{x_k\} \) satisfy the **first-order** necessary optimality condition.
  - notice that if \( g_k = 0 \), then Cauchy condition requires the trial step \( s_k \) to satisfy
    \[
    m_k(0) - m_k(s_k) \geq \frac{1}{2} \|g_k\| \min \left\{ \frac{\|g_k\|}{1 + \|B_k\|_2}, \delta_k \right\} = 0
    \]
    which is satisfied by \( s_k = 0 \).
  - thus, can’t prove that limit points of \( \{x_k\} \) for previous trust-region algorithms satisfy
    **second-order** necessary optimality conditions.
Question: Can we derive algorithms that guarantee that limit points satisfy second-order necessary conditions?

Answer: Yes, if we incorporate directions of negative curvature into the methods.

Roughly: iterates must be able to move away from a saddle point.

- **line search methods:**
  - should use directions of negative curvature (when they exist)
  - line search procedure should take these directions into account

- **trust-region methods:**
  - should use directions of negative curvature (when they exist)
  - should generalize Cauchy condition to account for negative curvature

- what about fixed step size methods?
Outline

1. Introduction

2. Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3. Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4. Trust-Region Methods

5. Conclusions and Final Thoughts
Outline

1. Introduction

2. Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3. Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4. Trust-Region Methods

5. Conclusions and Final Thoughts
The algorithm described here uses the update formula

\[ x_{k+1} = x_k + \alpha p_k \]  

(2)

where

\( \alpha > 0 \) is a fixed step size.

\( p_k = -g_k \), i.e., \( p_k \) is the steepest descent direction at \( x_k \)

**Question:** Can we guarantee that \( f(x_{k+1}) < f(x_k) \)?

**Answer:** Yes, if we assume global Lipschitz continuity of \( \nabla f \) and \( \alpha \) is small enough.

**Lemma 1**

If \( \nabla f \) is Lipschitz continuous with constant \( L > 0 \) and \( \alpha \in (0, 2/L) \), then

\[ f(x_{k+1}) \leq f(x_k) - c(\alpha) \|g_k\|^2_2 \]

where the constant \( c(\alpha) := \alpha(1 - \alpha L/2) > 0 \).

**Proof:** It follows from Lipschitz continuity of \( \nabla f \), (2), and \( p_k = -g_k \) that

\[
\begin{align*}
  f(x_{k+1}) &= f(x_k + \alpha p_k) \\
  &\leq f(x_k) + \alpha g_k^T p_k + (L/2)\alpha^2 \|p_k\|^2_2 \\
  &= f(x_k) - \alpha \|g_k\|^2_2 + (L/2)\alpha^2 \|g_k\|^2_2 \\
  &= f(x_k) - \alpha(1 - \alpha L/2) \|g_k\|^2_2 \\
  &\equiv f(x_k) - c(\alpha) \|g_k\|^2_2
\end{align*}
\]

which completes the proof. \( \blacksquare \)
The previous lemma suggest the following algorithm.

Algorithm 1 Fixed Step Size Algorithm

**Require:** let $L > 0$ be the Lipschitz constant for $\nabla f$.

1. Choose $\alpha \in (0, 2/L)$.
2. Choose $x_0 \in \mathbb{R}^n$ and set $k \leftarrow 0$.
3. loop
4. if $g_k = 0$ then
5. Return first-order solution $x_k$
6. end if
7. Set $p_k \leftarrow -g_k$.
8. Set $x_{k+1} \leftarrow x_k + \alpha p_k$.
9. Set $k \leftarrow k + 1$.
10. end loop

Comments

- In practice, should include a stopping condition and a maximum allowed iterations.
- A reasonable stopping condition is

$$\| \nabla f(x_k) \| \leq 10^{-5} \max\{1, \| \nabla f(x_0) \|_2\}$$
Theorem 2

Assume that \( \nabla f \) is Lipschitz continuous with Lipschitz constant \( L \) and \( f \) is bounded below, i.e., \( f(x) \geq f^* \) for all \( x \in \mathbb{R}^n \). Let \( \{x_k\} \) be the iterate sequence generated by Algorithm 1. Then for all \( T \geq 1 \),

\[
\min_{k=0,\ldots,T} \|\nabla f(x_k)\| \leq \sqrt{\frac{f(x_0) - f^*}{c(\alpha)(T + 1)}}.
\]

(4)

Proof: Summing (3) for \( k = 0, \ldots, T \), we obtain

\[
f^* \leq f(x_{T+1}) \leq f(x_0) - c(\alpha) \sum_{k=0}^{T} \|\nabla f(x_k)\|^2 \leq f(x_0) - c(\alpha)(T + 1) \min_{k=0,\ldots,T} \|\nabla f(x_k)\|^2
\]

Rearranging and taking square roots gives us (4).
Theorem 2 applies to any $f$ with Lipschitz continuous gradient. Can do better analysis more if $f$ is also convex and guarantee near global optimality.

Advantages of using a fixed step size:

- The gradient descent method with a fixed step size is very simple.
- It does not require "extra" evaluations of $f$ and $\nabla f$ as needed for linesearch methods.

Disadvantages of using a fixed step size:

- It appears that we need to know the value of the Lipschitz constant $L$. Not quite!
  - Note that in (3), it holds that if $\alpha \in (0, 1/L]$, then
    $$f(x_{k+1}) \leq f(x_k) - \frac{1}{2} \alpha \|g_k\|^2_2.$$  \hspace{1cm} (5)
  - During each iteration check if (5) holds, and if it doesn’t, then decrease $\alpha$ and try again, e.g., $\alpha \leftarrow \frac{1}{2} \alpha$.
  - Eventually, $\alpha$ will fall into the interval $(0, 1/L]$ as needed, and then (3) will always hold.
  - Seem sort of familiar? Similar, but not the same, as a linesearch.
- Even if we know $L$, it is a global Lipschitz constant
  - If $L$ is very large, then the step size $\alpha \in (0, 2/L)$, which likely will lead to slow convergence.
  - Ideally, you would want a local Lipschitz constant at each iterate $x_k$, i.e., we should choose $\alpha \in (0, 2/L_k)$ where $L_k$ denotes the local Lipschitz constant at $x_k$.
  - How to use search directions beyond steepest descent? Accepting the unit step?

An "optimal" $\alpha^*$ maximizes $c(\alpha)$ over $\alpha \in (0, 2/L)$, i.e., $\alpha^* := 1/L$.

- This "optimal" $\alpha^*$ usually does not perform the best because its calculation is based on the global Lipschitz constant, which usually overestimates local Lipschitz constants.
- Thus, the step size is often shorter than is actually "optimal".
- In practice, a step size $\alpha \in (0, 2/L)$ that is closer to $2/L$ often performs better.
Outline

1 Introduction

2 Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3 Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4 Trust-Region Methods

5 Conclusions and Final Thoughts
Second-order methods use second-derivatives. Since they use second-derivatives, ideally should aim for second-order optimality. Second-order necessary conditions:

\[ \nabla f(x^*) = 0 \quad \text{and} \quad \nabla^2 f(x^*) \succeq 0 \]

Therefore, anytime at \( x_k \) it holds that \( H_k \not\succeq 0 \), we should be prepared to compute a direction of “sufficient negative curvature”, i.e., a direction \( d_k \) satisfying

\[ d_k^T H_k d_k \leq \gamma \lambda_k \|d_k\|_2^2 \quad \text{and} \quad g_k^T d_k \leq 0 \] (6)

for some \( \gamma \in (0, 1] \) and where \( \lambda_k < 0 \) denotes the left-most eigenvalue of \( H_k \).

This definition ensures that \( d_k \) is a descent direction for \( f \) at \( x_k \), in the sense that there exists \( \epsilon > 0 \) such that \( f(x_k + \alpha d_k) < f(x_k) \) for all \( \alpha \in (0, \epsilon) \).

To get concrete convergence rates, we assume that a descent step \( p_k \) satisfying

\[ -g_k^T p_k \geq \delta \|p_k\|_2 \|g_k\|_2 \] (7)

is computed during iteration \( k \), where \( \delta \in (0, 1] \).
Examples for computing $p_k$ and $d_k$.

Examples for computing the descent direction $p_k$ satisfying (7):

- $p_k = -\tau g_k$ for any $\tau > 0$ since
  \[ -g_k^T p_k = \tau \|g_k\|^2 \geq \delta \tau \|g_k\|^2 = \delta \|\tau g_k\|_2 \|g_k\|_2 = \delta \|p_k\|_2 \|g_k\|_2 \]

- $B_k p_k = -g_k$ for any $B_k$ that is symmetric, positive-definite, and satisfies
  \[ \frac{\lambda_{\min}(B_k)}{\lambda_{\max}(B_k)} \geq \delta \quad \text{(exercise)} \]

Examples for computing the negative curvature direction $d_k$ satisfying (6):

- With $(\lambda_k, v_k)$ denoting a left-most eigenvalue-eigenvector pair for $H_k$, choose
  \[ d_k = \begin{cases} 
  v_k & \text{if } g_k^T v_k \leq 0 \\
  -v_k & \text{otherwise} 
  \end{cases} \]
  since
  \[ d_k^T H_k d_k = v_k^T H_k v_k = \lambda_k \|v_k\|^2 \leq \gamma \lambda_k \|v_k\|^2 = \gamma \lambda_k \|d_k\|^2 \]

- Compute $d_k$ via matrix-free Lanczos iterations (see [3]).
Motivation for an algorithm

- Let \( \{ L, \sigma \} \subset (0, \infty) \) be Lipschitz constants for \( \{ \nabla f, \nabla^2 f \} \), respectively.
- For \( L_k \in (0, \infty) \) and \( \sigma_k \in (0, \infty) \) we define the model reductions

\[
\begin{align*}
    m_{p,k}(\alpha) &:= -\alpha g_k^T p_k - \frac{1}{2} L_k \alpha^2 \|p_k\|_2^2 \\
    m_{d,k}(\beta) &:= -\beta g_k^T d_k - \frac{1}{2} \beta^2 d_k^T H_k d_k - \frac{\sigma_k}{6} \beta^3 \|d_k\|_2^3
\end{align*}
\]

because we know that if \( L_k \geq L \) and \( \sigma_k \geq \sigma \), then

\[
\begin{align*}
    f(x_k + \alpha p_k) &\leq f(x_k) - m_{p,k}(\alpha) \\
    f(x_k + \beta d_k) &\leq f(x_k) - m_{d,k}(\beta)
\end{align*}
\]

for all \( \alpha \in (0, \infty) \) and \( \beta \in (0, \infty) \).

- These inequalities suggest that we should choose

\[
\alpha_k = \arg\max_{\alpha \geq 0} m_{p,k}(\alpha) \quad \text{and} \quad \beta_k = \arg\max_{\beta \geq 0} m_{d,k}(\beta)
\]

which can be shown (exercise) to be given by

\[
\begin{align*}
    \alpha_k &:= -\frac{g_k^T p_k}{L_k \|p_k\|_2^2} \\
    \beta_k &:= \frac{\left(-c_k + \sqrt{c_k^2 - 2\sigma_k \|d_k\|_2^3 g_k^T d_k}\right)}{\sigma_k \|d_k\|_2^3}
\end{align*}
\]

where \( c_k := d_k^T H_k d_k \) is the curvature along the direction \( d_k \).

- The fact that \( p_k \) and \( d_k \) are descent directions implies that \( \alpha_k, \beta_k > 0 \).

- The method below chooses which step to use based on a smart test.
Algorithm 2 Dynamic Method

Require: \( \rho \in (1, \infty) \) and initial estimates \( \{L_1, \sigma_1\} \subset (0, \infty) \)

1: for \( k \in \mathbb{N} \) do
2: \hspace{1em} if \( \lambda_k \geq 0 \) then set \( d_k \leftarrow 0 \) else choose \( d_k \) satisfying (6)
3: \hspace{1em} if \( g_k = 0 \) then set \( p_k \leftarrow 0 \) else choose \( p_k \) satisfying (7)
4: \hspace{1em} if \( d_k = p_k = 0 \) then return \( x_k \)
5: loop
6: \hspace{1em} compute \( \alpha_k \geq 0 \) and \( \beta_k \geq 0 \) from (9)
7: \hspace{1em} if \( m_{p,k}(\alpha_k) \geq m_{d,k}(\beta_k) \) then
8: \hspace{2em} if (8a) holds then
9: \hspace{3em} set \( x_{k+1} \leftarrow x_k + \alpha_k p_k \) and then exit loop
10: \hspace{2em} else \hspace{1em} set \( L_k \leftarrow \rho L_k \)
11: \hspace{2em} end if
12: \hspace{1em} else
13: \hspace{2em} if (8b) holds then
14: \hspace{3em} set \( x_{k+1} \leftarrow x_k + \beta_k d_k \) and then exit loop
15: \hspace{2em} else \hspace{1em} set \( \sigma_k \leftarrow \rho \sigma_k \)
16: \hspace{2em} end if
17: \hspace{1em} end if
18: \hspace{1em} end if
19: \hspace{1em} end loop
20: \hspace{1em} set \( L_{k+1} \in (0, L_k] \) and \( \sigma_{k+1} \in (0, \sigma_k] \)
21: \hspace{1em} end for
Lemma 3 (well-defined)

Algorithm 2 either terminates finitely in Step 4 or generates infinitely many iterates. In addition, at the end of each iteration $k$, it holds that

\[ L_k \leq L_{\max} := \max\{L_1, \rho L\} \quad \text{and} \quad \sigma_k \leq \sigma_{\max} := \max\{\sigma_1, \rho \sigma\}. \quad (10) \]

Proof:
During iteration $k$, Algorithm 2 might finitely terminate. If it does not, then it enters the loop. If that loop were never to terminate, then the updates to $L_k$ and $\sigma_k$ would cause at least one of them to become arbitrarily large during that iteration inside the loop. Since (8a) holds whenever $L_k \geq L$, and (8b) holds whenever $\sigma_k \geq \sigma$, it follows that the loop must eventually terminate, thus proving the result.

The fact that (10) holds follows from the fact that $L_k$ and $\sigma_k$ are increased by the factor $\rho$ and, as stated above, may only be needed when $L_k < L$ and $\sigma_k < \sigma$. \qed
Theorem 4 (limit points are second-order points (see [2, Theorem 2]))

When Algorithm 2 is executed, one of the following must hold:

(i) Algorithm 2 terminates in Step 4 with $g_k = 0$ and $\lambda_k \geq 0$, i.e., $x_k$ satisfies the 2nd-order necessary optimality conditions.

(ii) The objective function is unbounded below over the sequence of iterates, i.e.,

$$\lim_{k \to \infty} f(x_k) = -\infty$$

(iii) Infinitely many iterates are computed and they satisfy

$$\lim_{k \to \infty} \|g_k\|_2 = 0 \quad \text{and} \quad \liminf_{k \to \infty} \lambda_k \geq 0$$

so that all limit points of $\{x_k\}$ satisfy the 2nd-order necessary optimality conditions.

Proof:
Algorithm 2 terminates finitely if and only if $d_k = p_k = 0$ for some $k$. This can only occur if $\lambda_k \geq 0$ and $g_k = 0$, which is the outcome for part (i). Thus, for the remainder of the proof, we assume that Algorithm 2 does not terminate finitely.

Moreover, we will also assume that case (ii) does not happen and prove that case (iii) must occur, i.e., we assume that there a scalar $f_{low} > -\infty$ such that

$$f(x_k) \geq f_{low} \quad \text{for all } k$$

and proceed to prove that case (iii) must happen.
If $p_k \neq 0$, then the definition of $\alpha_k$ in (9) and the condition on $p_k$ in (7) ensure that

$$
m_{p,k}(\alpha_k) = -\alpha_k g_k^T p_k - \frac{1}{2} L_k \alpha_k^2 \|p_k\|_2^2
$$

$$
= - \left( \frac{-g_k^T p_k}{L_k \|p_k\|_2^2} \right) (g_k^T p_k) - \frac{1}{2} L_k \left( \frac{-g_k^T p_k}{L_k \|p_k\|_2^2} \right)^2 \|p_k\|_2^2
$$

$$
= \frac{1}{L_k} \left( \frac{g_k^T p_k}{\|p_k\|_2} \right)^2 - \frac{1}{2L_k} \left( \frac{g_k^T p_k}{\|p_k\|_2} \right)^2
$$

$$
= \frac{1}{2L_k} \left( \frac{g_k^T p_k}{\|p_k\|_2} \right)^2 \geq \frac{\delta^2}{2L_k} \|g_k\|_2^2. 
$$

(13)

Similarly, if $d_k \neq 0$, then since $\beta_k$ maximizes $m_{d,k}(\beta)$ over $\beta > 0$, defining

$$
\beta_k := -2d_k^T H_k d_k / (\sigma_k \|d_k\|_2^3) > 0
$$

and using (6) we find that

$$
m_{d,k}(\beta_k) \geq m_{d,k}(\hat{\beta}_k)
$$

$$
= -\hat{\beta}_k g_k^T d_k - \frac{1}{2} \hat{\beta}_k^2 d_k^T H_k d_k - \frac{1}{6} \sigma_k \hat{\beta}_k^3 \|d_k\|_2^3
$$

$$
\geq -\frac{1}{2} \hat{\beta}_k^2 d_k^T H_k d_k - \frac{1}{6} \sigma_k \hat{\beta}_k^3 \|d_k\|_2^3
$$

$$
= -\frac{2(d_k^T H_k d_k)^3}{3\sigma_k^2 \|d_k\|_2^6} \geq \frac{2\gamma^3}{3\sigma_k^2} |\lambda_k|^3.
$$

(14)
\[ \beta \geq 0 \]
\[ m_{d,k}(\beta) = -\beta g^T \mathbf{d}_k - \frac{\beta^2 d_k H_k d_k}{2} - \frac{\sigma_k}{6} \mathbf{B}^3 ||d_k||^3 \]
\[ g^T \mathbf{d}_k \leq 0 \]
\[ \mathbf{d}_k \approx -\beta d_k H_k d_k - \frac{\sigma_k}{2} \beta^2 ||d_k||^3 \]
\[ \tilde{m}_{d,k}(\beta) = -\beta d_k H_k d_k - \frac{\sigma_k}{2} \beta^2 ||d_k||^3 \]
\[ \maximize \tilde{m}_{d,k}(\beta) \]
\[ \tilde{m}_{d,k}(\beta) = -\beta d_k H_k d_k - \frac{\sigma_k}{2} \beta^2 ||d_k||^3 \]
\[ = 0 \]
We now claim that, for all $k$, it holds that
\[
f(x_k) - f(x_{k+1}) \geq \max\left\{ \frac{\delta^2}{2L_k} \|g_k\|_2^2, \frac{2\gamma^3}{3\sigma_k^2} \left| \min(0, \lambda_k) \right|^3 \right\},
\]
which we prove by considering two cases.

**Case 1:** the update $x_{k+1} \leftarrow x_k + \alpha_k p_k$ is completed.
It must be that $p_k \neq 0$, (8a), and $m_{p,k}(\alpha_k) \geq m_{d,k}(\beta_k)$ hold (see Algorithm 2).
Combining these facts with (13) and (14) shows that
\[
f(x_k + \alpha_k p_k) \leq f(x_k) - m_{p,k}(\alpha_k)
= f(x_k) - \max\{m_{p,k}(\alpha_k), m_{d,k}(\beta_k)\}
\leq f(x_k) - \max\left\{ \frac{\delta^2}{2L_k} \|g_k\|_2^2, \frac{2\gamma^3}{3\sigma_k^2} \left| \min(0, \lambda_k) \right|^3 \right\}
\]
which gives (15); note the last inequality holds regardless of whether $d_k$ is nonzero.

**Case 2:** the update $x_{k+1} \leftarrow x_k + \beta_k d_k$ is completed.
It must be that $d_k \neq 0$, (8b), and $m_{p,k}(\alpha_k) < m_{d,k}(\beta_k)$ hold (see Algorithm 2).
Combining these facts with (13) and (14) show that
\[
f(x_k + \beta_k d_k) \leq f(x_k) - m_{d,k}(\beta_k)
= f(x_k) - \max\{m_{p,k}(\alpha_k), m_{d,k}(\beta_k)\}
\leq f(x_k) - \max\left\{ \frac{\delta^2}{2L_k} \|g_k\|_2^2, \frac{2\gamma^3}{3\sigma_k^2} \left| \min(0, \lambda_k) \right|^3 \right\}
\]
which gives (15); note the last inequality holds regardless of whether $p_k$ is nonzero.
For each $\ell \in \mathbb{N}$, it follows from (15) and the bounds in (10) that

$$f_1 - f_{\ell+1} = \sum_{k=1}^{\ell} (f(x_k) - f(x_{k+1}))$$

$$\geq \sum_{k=1}^{\ell} \max \left\{ \frac{\delta^2}{2L_k} \|g_k\|_2^2, \frac{2\gamma^3}{3\sigma_k^2} |\min(0, \lambda_k)|^3 \right\}$$

$$\geq \sum_{k=1}^{\ell} \max \left\{ \frac{\delta^2}{2L_{\max}} \|g_k\|_2^2, \frac{2\gamma^3}{3\sigma_{\max}^2} |\min(0, \lambda_k)|^3 \right\}$$

from which it follows, by letting $\ell \to \infty$ and using (12), that

$$\infty > f_0 - f_{\text{low}} \geq \sum_{k=1}^{\infty} \max \left\{ \frac{\delta^2}{2L_{\max}} \|g_k\|_2^2, \frac{2\gamma^3}{3\sigma_{\max}^2} |\min(0, \lambda_k)|^3 \right\}.$$ 

It now trivially follows that

$$\min_{\mathcal{D}_T} \sum_{k=1}^{\infty} \|g_k\|_2^2 < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} |\min(0, \lambda_k)| < \infty$$

which implies that

$$\lim_{k \to \infty} \|g_k\|_2 = 0 \quad \text{and} \quad \lim_{k \to \infty} |\min(0, \lambda_k)| = 0.$$ 

The proof is finished because the previous limits are equivalent to (11).
The convergence theory allows $L_{k+1} \leftarrow L_k$ and $\sigma_{k+1} \leftarrow \sigma_k$ in Step 21.

- In this case, the Lipschitz constant estimates are monotonically increasing.
- In Algorithm 2 we allow them to decrease since this might yield better practical results.

What if the Lipschitz constants $L$ and $\sigma$ for $g$ and $H$, respectively, are known?

- One could set $L_k = L$ and $\sigma_k = \sigma$ for each iteration, so that the loop is not needed.
- This would generally lead to more iterations (when a termination condition is used).
- If the cost of evaluating $f$ is substantial, this fixed parameter choice might work well.

Each time through the loop, condition (8a) or (8b) is tested, but not both.

- Evaluating both would require an extra evaluation of $f$.
- If the cost of evaluating $f$ is not a concern, then one could choose between the $p_k$ and $d_k$ based on the actual objective function decrease rather than on the model decrease.

Can also adapt to get standard convergence rate: after $O \left( \frac{1}{\epsilon}^3 \right)$ steps, we must have visited a point where the gradient norm is at most $\epsilon$ and the minimum eigenvalue of the Hessian are at least $-\epsilon$. 
1 Introduction

2 Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3 Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4 Trust-Region Methods

5 Conclusions and Final Thoughts
Outline

1 Introduction

2 Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3 Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4 Trust-Region Methods

5 Conclusions and Final Thoughts
Motivation:
- First-order linesearch algorithms use descent directions $p_k$ to prove
  \[
  \lim_{k \to \infty} \|g_k\|_2 = 0
  \]
  under commonly made assumptions on $f$ and the directions $\{p_k\}$.
- This is not sufficient for establishing second-order conditions.
- For example, consider the situation that $x_k$ is a saddle point:
  - It follows that $g_k = 0$ and $\lambda_{\text{min}}(H_k) < 0$.
  - There does not exist any direction $p_k$ satisfying $g_k^T p_k < 0$ (no descent direction).
  - First-order methods must terminate even if $\lambda_k \equiv \lambda_{\text{min}}(H_k) < 0$.
- Second-order methods must be allowed to use directions of negative curvature.

Basic Challenges:
- Develop linesearch methods that use descent and negative curvature directions.
- What properties should negative curvature directions satisfy?
- Generalize linesearch conditions (we will focus on the Armijo condition).
Outline

1 Introduction

2 Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3 Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4 Trust-Region Methods

5 Conclusions and Final Thoughts
Conditions required of a descent direction $p_k$

At iterate $x_k$, we should compute any descent direction $p_k$ satisfying

$$g_k^T p_k \leq -\gamma_p \|g_k\|_2 \|p_k\|_2 \quad \text{and} \quad (1/\kappa_p) \|g_k\|_2 \leq \|p_k\|_2 \leq \kappa_p \|g_k\|_2$$

for some chosen constants $\gamma_p \in (0, 1]$ and $\kappa_p \in [1, \infty)$.

Comments:

- Condition (16) ensures that $p_k$ is a “sufficient” descent direction.
- Condition (17) ensures that $p_k$ is roughly the same size as $g_k$.
- Both conditions hold in the following settings (exercises):
  - holds for $p_k = -g_k$
  - holds for $B_k p_k = -g_k$ provided $B_k$ is chosen to be symmetric, and there exists $\lambda_{\min}$ and $\lambda_{\max}$ satisfying

$$0 < \lambda_{\min} \leq \lambda_{\min}(B_k) \leq \lambda_{\max}(B_k) \leq \lambda_{\max} < \infty \quad \text{for all} \ k$$
Conditions required of a negative curvature direction $d_k$

At iterate $x_k$, we should compute any negative curvature direction $d_k$ satisfying

$$d_k^T H_k d_k \leq \gamma_d \lambda_k \|d_k\|_2^2$$  \hspace{1cm} (18)

$$(1/\kappa_d)|\min(0, \lambda_k)| \leq \|d_k\|_2 \leq \kappa_d |\min(0, \lambda_k)|$$ \hspace{1cm} and \hspace{1cm} (19)

$$d_k^T g_k \leq 0$$  \hspace{1cm} (20)

for some constants $\gamma_d \in (0, 1]$ and $\kappa_d \in [1, \infty)$.

Comments:

- Condition (18) ensures that $d_k$ is a direction of negative curvature. Moreover, since

$$\frac{d_k^T H_k d_k}{\|d_k\|_2^2} \leq \gamma_d \lambda_k = \gamma_d \frac{v_k^T H_k v_k}{\|v_k\|_2^2}$$

we know that $d_k$ achieves a fraction ($\gamma_d$) of the most negative curvature.

- Condition (19) ensures that $d_k$ is roughly the size of $|\lambda_k|$, when $\lambda_k < 0$.

- Condition (20) ensures that $d_k$ is a non-ascent direction.

- All three conditions are satisfied in the following settings (exercises):
  - holds for $d_k = \pm |\min(0, \lambda_k)| v_k$ (the sign $\pm$ is chosen to ensure that (20) holds)
  - holds for any sufficiently accurate (that depends on the choice of $\gamma_d$) approximation to the $v_k$ in the previous bullet point.
How do we use $p_k$ and $d_k$?

**Definition 5 (weighting function)**

We say that a function $\omega : \mathbb{R} \rightarrow [0, \infty)$ is a **weighting function** if and only if

1. $\omega(\cdot)$ is twice continuously differentiable
2. $\omega(0) = 0$

**A curvilinear search path**

Given $x_k$, $p_k$, $d_k$, and weighting functions $\omega_{k,p}$ and $\omega_{k,d}$, the **curvilinear search path** is

$$\{x_k + s_k(\alpha) \in \mathbb{R}^n \mid \alpha \geq 0\} \subset \mathbb{R}^n$$

where

$$s_k(\alpha) := \omega_{k,p}(\alpha)p_k + \omega_{k,d}(\alpha)d_k$$

**Comments:**

- $s_k(\cdot)$ is twice continuously differentiable because $w_{k,p}(\cdot)$ and $w_{k,d}(\cdot)$ are.
- The most popular choice for weighting functions is

$$w_{k,p}(\alpha) = w_{k,d}(\alpha) = \alpha$$

so that

$$s_k(\alpha) = \alpha(p_k + d_k)$$
Curvilinear search path for $\alpha \in [0, 1]$ (purple line segment) for the case

$$w_{k,p}(\alpha) = w_{k,d}(\alpha) = \alpha$$
Curvilinear search path for $\alpha \in [0, 1]$ (purple line segment) for the case

$$w_{k,p}(\alpha) = \alpha \quad \text{and} \quad w_{k,d}(\alpha) = \alpha^2$$
A second-order “linesearch”

Define $\phi_k : \mathbb{R} \to \mathbb{R}$ as the objective function evaluated on the curvilinear path:

$$\phi_k(\alpha) := f(x_k + s_k(\alpha))$$

so that we have

$$\phi_k'(\alpha) = \nabla f(x_k + s_k(\alpha))^T s_k'(\alpha) \quad (21)$$

$$\phi_k''(\alpha) = s_k'(\alpha)^T \nabla^2 f(x_k + s_k(\alpha)) s_k'(\alpha) + \nabla f(x_k + s_k(\alpha))^T s_k''(\alpha) \quad (22)$$

which may be combined with $s_k(0) = 0$ (why?) to obtain

$$\phi_k'(0) = g_k^T s_k'(0) \quad (23)$$

$$\phi_k''(0) = s_k'(0)^T H_k s_k'(0) + g_k^T s_k''(0) \quad (24)$$

Definition 6 (second-order Armijo condition)

Given $\eta \in (0, 1)$, we say that $\bar{\alpha} > 0$ satisfies the second-order Armijo condition if

$$\phi_k(0) - \phi_k(\bar{\alpha}) \geq \eta (q_k(0) - q_k(\bar{\alpha})) \quad (25)$$

where the (potentially) quadratic function $q_k$ is defined as

$$q_k(\alpha) := \phi_k(0) + \phi_k'(0) \alpha + \frac{1}{2} \min\{0, \phi_k''(0)\} \alpha^2$$
Lemma 7 (2nd-order Armijo condition is satisfied for all $\alpha$ sufficiently small)

If either
\[ \phi_k'(0) < 0, \quad \text{or} \quad \phi_k'(0) \leq 0 \quad \text{and} \quad \phi_k''(0) < 0 \]  
(26)

then the second-order Armijo condition (25) holds for all sufficiently small $\alpha > 0$.

Proof:
Motivated by the second-order Armijo condition, let us define the function
\[ \psi_k(\alpha) := \phi_k(0) - \phi_k(\alpha) - \eta(q_k(0) - q_k(\alpha)). \]

Note that
\[ \psi_k(0) = 0 \]
and that
\[ \psi_k'(\alpha) = -\phi_k'(\alpha) + \eta q_k'(\alpha) \]
\[ \psi_k''(\alpha) = -\phi_k''(\alpha) + \eta q_k''(\alpha) \]

so that in particular
\[ \psi_k'(0) = -\phi_k'(0) + \eta \phi_k'(0) = (\eta - 1)\phi_k'(0) \]
(27)
\[ \psi_k''(0) = -\phi_k''(0) + \eta \min\{0, \phi_k''(0)\} = \min\{-\phi_k''(0), (\eta - 1)\phi_k''(0)\}. \]
(28)

Now consider the two cases in the statement of the lemma.
Case 1: $\phi_k'(0) < 0$.

In this case, we have from (27) and $\eta \in (0, 1)$ that

$$\psi_k'(0) > 0$$

so that $\psi_k(\alpha) > 0$ for all sufficiently small $\alpha > 0$, which is equivalent to (25).

Case 2: $\phi_k'(0) \leq 0$ and $\phi_k''(0) < 0$.

In this case, we have from (27), (28), and $\eta \in (0, 1)$ that

$$\psi_k'(0) \geq 0 \quad \text{and} \quad \psi_k''(0) > 0$$

so that $\psi_k(\alpha) > 0$ for all sufficiently small $\alpha > 0$, which is equivalent to (25).
Lemma 8

If the following hold:
- $p_k$ is computed to satisfy (16)–(17)
- $d_k$ is computed to satisfy (18)–(20)
- at least one of $p_k$ and $d_k$ is nonzero, and
- we choose $\omega_{k,p}(\alpha) = \omega_{k,d}(\alpha) = \alpha$
then (26) holds.

Proof:
Note that from the choice of the weight functions, we have

\[ s_k(\alpha) = \alpha p_k + \alpha d_k = \alpha (p_k + d_k) \]

so that

\[ s_k'(0) = p_k + d_k \quad \text{and} \quad s_k''(0) = 0. \]

Combining this with (23) and (24) it follows that

\[ \phi_k'(0) = g_k^T (p_k + d_k) \quad \text{and} \quad \phi_k''(0) = (p_k + d_k)^T H_k (p_k + d_k). \]  

(29)

We now consider two cases.
Case 1: $p_k \neq 0$

In this case, it follows from (17) that $g_k \neq 0$ and then from (16) that $g_k^T p_k \leq -\gamma_p \|g_k\|_2 \|p_k\|_2 < 0$. Combining this fact with (20) and (29) shows that

$$
\phi'_k(0) = g_k^T (p_k + d_k) = g_k^T p_k + g_k^T d_k \leq g_k^T p_k < 0
$$

which establishes that (26) holds in this case.

Case 2: $p_k = 0$

In this case, it follows from (20) that

$$
\phi'_k(0) = g_k^T (p_k + d_k) = g_k^T d_k \leq 0,
$$

which shows that the first condition in the second possible scenario of (26) holds.

Next, let us observe from the statement of the lemma that $d_k \neq 0$ (because $p_k = 0$ in this case), which by (19) also ensures that $\lambda_k < 0$. By combining these facts with (29), $p_k = 0$, and (18) we obtain

$$
\phi''_k(0) = (p_k + d_k)^T H_k (p_k + d_k) = d_k^T H_k d_k \leq \gamma d \lambda_k \|d_k\|_2^2 < 0
$$

which shows that the second condition in the second possible scenario of (26) holds.

Remark: By combining Lemma 7 and Lemma 8 we know that the following general curvilinear search method is well defined.
Algorithm 3 General curvilinear search method

Require: \( \{\tau, \eta\} \subset (0, 1) \) and \( \alpha_{init} \in (0, \infty) \).
1: Choose \( x_0 \) and evaluate \( f_0, g_0 \), and \( H_0 \).
2: for \( k \in \{0, 1, 2, \ldots\} \) do
3: Compute a descent direction \( p_k \) satisfying (16)–(17).
4: Compute a negative curvature direction \( d_k \) satisfying (18)–(20).
5: Choose weight functions \( \omega_{k,p}(\cdot) \) and \( \omega_{k,d}(\cdot) \) so that (26) holds.
6: Set \( \alpha_k \leftarrow \alpha_{init} \)
7: loop \( \triangleright \) curvilinear search
8: if \( (\phi_k(0) - \phi_k(\alpha_k)) \geq \eta(q_k(0) - q_k(\alpha_k)) \) then
9: exit loop with the current \( \alpha_k \) value \( \triangleright \) 2nd-order Armijo
10: else
11: Set \( \alpha_k \leftarrow \tau \alpha_k \)
12: end if
13: end loop
14: Set \( x_{k+1} \leftarrow x_k + s_k(\alpha_k) \)
15: Evaluate \( f(x_{k+1}), g_{k+1}, \) and \( H_{k+1} \)
16: end for

Comments:
- A common choice is \( \alpha_{init} = 1 \) when Newton-like descent directions are computed.
- A reasonable termination condition should be used, such as

\[
\|g_k\| \leq 10^{-6} \max\{1, \|g_0\|\} \quad \text{and} \quad \lambda_k \geq 10^{-6} \min\{-1, \lambda_0\}
\]
1. Introduction

2. Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3. Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4. Trust-Region Methods

5. Conclusions and Final Thoughts
Our analysis makes use of the iteration index set

\[ S := \{ k : \phi''_k(0) < 0 \} \]  \hspace{1cm} (30)

which are the iterations that a true second-order model is used in the Armijo search.

We will also use the assumption that \( f \) is bounded below on the initial level set.

**Assumption 3.1 (bounded objective on the initial level set)**

The objective function \( f \) is bounded below on the level set

\[ L_0 := \{ x \in \mathbb{R}^n : f(x) \leq f(x_0) \} \]

We also use the following additional notation:

- \( L_{k,0} \) : denotes the Lipschitz constant for \( \phi'_k(\alpha) \) at \( \alpha = 0 \)
- \( \sigma_{k,0} \) : denotes the Lipschitz constant for \( \phi''_k(\alpha) \) at \( \alpha = 0 \)
Lemma 9

If \( k \notin S \), then

\[
\phi_k''(0) \geq 0 \quad \text{and} \quad \alpha_k \geq \min \left\{ \alpha_{\text{init}}, \frac{2\tau(1-\eta)|\phi'_k(0)|}{L_{k,0}} \right\}
\]

Proof: Since \( k \notin S \) it follows that \( \phi_k''(0) \geq 0 \), which proves (31). From Step 5 of Algorithm 3 (specifically that (26) holds) we have \( \phi'_k(0) < 0 \). Using this fact and a Taylor’s approximation shows that if

\[
\alpha \in \left(0, \frac{2(\eta - 1)\phi'_k(0)}{L_{k,0}}\right]
\]

then

\[
\phi_k(\alpha) \leq \phi_k(0) + \phi_k(0)'\alpha + \frac{1}{2}L_{k,0}\alpha^2 \\
\leq \phi_k(0) + \phi'_k(0)\alpha + (\eta - 1)\phi'_k(0)\alpha = \phi_k(0) + \eta\phi'_k(0)\alpha,
\]

so that the Armijo condition (25) holds for \( \alpha \). It follows from this fact and the strategy for updating \( \alpha_k \) in Algorithm 3 that

\[
\alpha_k \geq \min \left\{ \alpha_{\text{init}}, \frac{2\tau(1-\eta)|\phi'_k(0)|}{L_{k,0}} \right\}
\]

as claimed in (32), which completes the proof.  \( \square \)
Lemma 10

If Assumption 3.1 holds and the complement of the set $S$, i.e., $\mathbb{N} \setminus S$ has infinite cardinality, then

$$\lim \min_{k \notin S} \left\{ |\phi_k'(0)|, \frac{|\phi_k'(0)|^2}{L_{k,0}} \right\} = 0$$

Proof: Consider any $k \notin S$. It then follows from (31) and the fact that $\alpha_k$ satisfies the Armijo condition in Step 8 that

$$f(x_{k+1}) = f(x_k + s(\alpha_k))$$
$$= \phi_k(\alpha_k)$$
$$\leq \phi_k(0) - \eta(q_k(0) - q_k(\alpha_k))$$
$$= \phi_k(0) + \eta \phi_k'(0) \alpha_k = f(x_k) - \eta |\phi_k'(0)| \alpha_k.$$

It follows from this inequality, Assumption 3.1, and a standard argument that

$$\lim_{k \notin S} |\phi_k'(0)| \alpha_k = 0. \quad (33)$$

Combining this limit with (32) gives

$$\lim \min_{k \notin S} \left\{ \alpha_{\text{init}} |\phi_k'(0)|, \frac{2\tau (1 - \eta)|\phi_k'(0)|^2}{L_{k,0}} \right\} = 0.$$

Since $\eta \in (0, 1)$ and $\alpha_{\text{init}} \in (0, \infty)$, we can conclude that

$$\lim \min_{k \notin S} \left\{ |\phi_k'(0)|, \frac{|\phi_k'(0)|^2}{L_{k,0}} \right\} = 0.$$
Lemma 11

If \( k \in S \), then

\[
\alpha_k \geq \min \left\{ \alpha_{\text{init}}, \frac{3\tau (1-\eta) |\phi_k''(0)|}{\sigma_{k,0}} \right\}
\]

\[
\alpha_k \geq \min \left\{ \alpha_{\text{init}}, \frac{2\tau (1-\eta) |\phi_k'(0)|}{L_{k,0} + \eta |\phi_k''(0)|} \right\}
\]

(34)

Proof: We break the proof into two scenarios based on (34).

**Scenario 1:** Suppose that \( \alpha \) satisfies

\[
\alpha \in \left(0, \frac{3(\eta - 1) \phi_k''(0)}{\sigma_{k,0}} \right].
\]

(35)

Then, it follows from a Taylor’s approximation, \( \phi_k'(0) \leq 0 \) (see (26)), and \( \eta \in (0, 1) \) that

\[
\phi_k(\alpha) \leq \phi_k(0) + \phi_k'(0) \alpha + \frac{1}{2} \phi_k''(0) \alpha^2 + \frac{\sigma_{k,0}}{6} \alpha^3
\]

\[
\leq \phi_k(0) + \eta \phi_k'(0) \alpha + \frac{1}{2} \phi_k''(0) \alpha^2 + \frac{1}{2} (\eta - 1) \phi_k''(0) \alpha^2
\]

\[
= \phi_k(0) + \eta \left( \phi_k'(0) \alpha + \frac{1}{2} \phi_k''(0) \alpha^2 \right)
\]

\[
= \phi_k(0) - \eta \left( q_k(0) - q_k(\alpha) \right)
\]

meaning that the second-order Armijo condition (25) holds for \( \alpha \).
Scenario 2: Suppose that $\alpha$ satisfies

$$\alpha \in \left(0, \frac{2(\eta - 1)\phi'_k(0)}{L_{k,0} - \eta\phi''_k(0)}\right].$$  \hspace{1cm} (36)$$

Then, it follows from a Taylor's approximation, $\phi'_k(0) < 0$ (see (36)), and $\eta \in (0, 1)$ that

$$\phi_k(\alpha) \leq \phi_k(0) + \phi_k(0)' \alpha + \frac{1}{2}L_{k,0}\alpha^2$$
$$= \phi_k(0) + \phi'_k(0)\alpha + \frac{1}{2}L_{k,0}\alpha^2 - \frac{1}{2}\eta\phi''_k(0)\alpha^2 + \frac{1}{2}\eta\phi''_k(0)\alpha^2$$
$$= \phi_k(0) + \phi'_k(0)\alpha + \frac{1}{2}(L_{k,0} - \eta\phi''_k(0))\alpha^2 + \frac{1}{2}\eta\phi''_k(0)\alpha^2$$
$$= \phi_k(0) + \phi'_k(0)\alpha + (\eta - 1)\phi'_k(0)\alpha + \frac{1}{2}\eta\phi''_k(0)\alpha^2$$
$$= \phi_k(0) + \eta(\phi'_k(0)\alpha + \frac{1}{2}\phi''_k(0)\alpha^2)$$
$$= \phi_k(0) - \eta(q_k(0) - q_k(\alpha))$$

meaning that the second-order Armijo condition (25) holds for $\alpha$.

It follows from both scenarios and the update strategy for $\alpha_k$ in Algorithm 3 that (34) holds, which completes the proof.  \qed
Lemma 12

If Assumption 3.1 holds and $\mathcal{S}$ has infinite cardinality, then

$$0 = \lim_{k \in \mathcal{S}} \min \left\{ |\phi_k'(0)|, \frac{|\phi_k'(0)|^2}{L_{k,0} + \eta |\phi_k''(0)|} \right\} \quad \text{and}$$

$$0 = \lim_{k \in \mathcal{S}} \min \left\{ |\phi_k''(0)|, \frac{|\phi_k''(0)|^3}{(\sigma_{k,0})^2} \right\}. \quad (37)$$

(38)

Proof:
Let $k \in \mathcal{S}$ so that $\phi_k''(0) < 0$ and the second-order Armijo condition gives

$$f(x_{k+1}) = f(x_k + s_k(\alpha_k))$$

$$= \phi_k(\alpha_k)$$

$$\leq \phi_k(0) - \eta (q_k(0) - q_k(\alpha_k))$$

$$= \phi_k(0) - \eta (-\phi_k'(0)\alpha_k - \frac{1}{2}\phi_k''(0)\alpha_k^2)$$

$$= f(x_k) - \eta |\phi_k'(0)|\alpha_k - \frac{1}{2}\eta |\phi_k''(0)|\alpha_k^2$$

where we also used $\phi_k'(0) \leq 0$ (see (26)). Combining this with Assumption 3.1 gives

$$\lim_{k \in \mathcal{S}} |\phi_k'(0)|\alpha_k = 0 \quad \text{and} \quad \lim_{k \in \mathcal{S}} |\phi_k''(0)|\alpha_k^2 = 0.$$
On the previous slide we proved that
\[ \lim_{k \in S} |\phi'_k(0)| \alpha_k = 0 \quad \text{and} \quad \lim_{k \in S} |\phi''_k(0)| \alpha_k^2 = 0. \] (39)

By combing the first limit in (39) with the second inequality in (34) we find that
\[ 0 = \lim_{k \in S} |\phi'_k(0)| \min \left\{ \alpha_{init}, \frac{2\tau(1-\eta)|\phi'_k(0)|}{L_k,0 + \eta|\phi''_k(0)|} \right\} \]
\[ = \lim_{k \in S} \min \left\{ \alpha_{init}|\phi'_k(0)|, \frac{2\tau(1-\eta)|\phi'_k(0)|^2}{L_k,0 + \eta|\phi''_k(0)|} \right\} \]
which gives
\[ \lim_{k \in S} \min \left\{ |\phi'_k(0)|, \frac{|\phi'_k(0)|^2}{L_k,0 + \eta|\phi''_k(0)|} \right\} = 0, \]
i.e., that (37) holds. On the other hand, the second limit in (39) and the first inequality in (34) together give
\[ 0 = \lim_{k \in S} |\phi''_k(0)| \min \left\{ \alpha_{init}, \frac{3\tau(1-\eta)|\phi''_k(0)|}{\sigma_k,0} \right\}^2 \]
\[ = \lim_{k \in S} \min \left\{ \alpha_{init}^2|\phi''_k(0)|, \frac{9\tau^2(1-\eta)^2|\phi''_k(0)|^3}{(\sigma_k,0)^2} \right\} \]
which means that
\[ \lim_{k \in S} \min \left\{ |\phi''_k(0)|, \frac{|\phi''_k(0)|^3}{(\sigma_k,0)^2} \right\} = 0 \]
thus establishing that (38) holds. This completes the proof. \[\Box\]
**Theorem 13**

*The iterates generated by Algorithm 3 satisfy*

\[
0 = \lim_{k \to \infty} \min \left\{ |\phi_k'(0)|, \frac{|\phi_k'(0)|^2}{L_{k,0} + \eta \min\{0, \phi_k''(0)\}} \right\} \quad \text{and} \quad (40)
\]

\[
0 = \lim_{k \to \infty} \min \left\{ |\min\{0, \phi_k''(0)\}|, \frac{|\min\{0, \phi_k''(0)\}|^3}{(\sigma_{k,0})^2} \right\} \quad (41)
\]

**Proof:** The result follows from Lemma 12, Lemma 10, (31), and the definition of \(S\). □

**Big Picture Comment:** The idea is to combine Theorem 13 with specific choices for the weighting functions to derive convergence results, namely, that

\[
\lim_{k \to \infty} \|g_k\|_2 = 0 \quad \text{and} \quad \liminf_{k \to \infty} \lambda_k \geq 0
\]
Outline

1 Introduction

2 Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3 Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
     - An algorithmic instance: equal weighting

4 Trust-Region Methods

5 Conclusions and Final Thoughts
Instance: $\omega_{k,p}(\alpha) = \omega_{k,d}(\alpha) = \alpha$ for all $k$

In this case we have

$$s_k(\alpha) = \alpha(p_k + d_k)$$
$$s'_k(\alpha) = p_k + d_k$$
$$s''_k(\alpha) = 0$$
$$\phi'_k(\alpha) = \nabla f(x_k + \alpha(p_k + d_k))^T(p_k + d_k)$$
$$\phi''_k(\alpha) = (p_k + d_k)^T \nabla^2 f(x_k + \alpha(p_k + d_k))(p_k + d_k)$$
$$\phi'_k(0) = g_k^T(p_k + d_k)$$
$$\phi''_k(0) = (p_k + d_k)^T H_k(p_k + d_k)$$

We also recall that Lemma 8 ensures that either

(i) during some iteration it so-happens that $p_k = d_k = 0$, in which case $x_k$ already satisfies the second-order necessary optimality conditions; or

(ii) the conditions in (26) will be satisfied as required by Step 5 of Algorithm 3, so that Algorithm 3 is well-posed for this choice.

For the remainder of this example, we assume that case (ii) occurs.
Instance: \( \omega_{k,p}(\alpha) = \omega_{k,d}(\alpha) = \alpha \) for all \( k \)

Next, we can observe by using the Lipschitz constant \( L \) for \( \nabla f \) that

\[
|\phi_k'(\alpha) - \phi_k'(0)| = |\nabla f(x_k + \alpha(p_k + d_k))^T(p_k + d_k) - g_k^T(p_k + d_k)|
\leq \|\nabla f(x_k + \alpha(p_k + d_k)) - g_k\|_2\|p_k + d_k\|_2
\leq (L\|p_k + d_k\|_2^2)\alpha \text{ for all } \alpha
\]

so that

\[
L_{k,0} \leq L\|p_k + d_k\|_2^2. \tag{42}
\]

Similarly, we can observe by using the Lipschitz constant \( \sigma \) for \( \nabla^2 f \) that

\[
|\phi_k''(\alpha) - \phi_k''(0)|
= |(p_k + d_k)^T\nabla^2 f(x_k + \alpha(p_k + d_k))(p_k + d_k) - (p_k + d_k)^T H_k(p_k + d_k)|
\leq \|\nabla^2 f(x_k + \alpha(p_k + d_k)) - H_k\|_2\|p_k + d_k\|_2^2
\leq \left(\sigma\|p_k + d_k\|_2^3\right)\alpha \text{ for all } \alpha
\]

so that

\[
\sigma_{k,0} \leq \sigma\|p_k + d_k\|_2^3. \tag{43}
\]
Instance: \(\omega_{k,p}(\alpha) = \omega_{k,d}(\alpha) = \alpha\) for all \(k\)

It now follows from (40), (42), (20), (16), (17), and the definition of \(\lambda_k\) that

\[
0 = \lim_{k \to \infty} \min_{\kappa \in \mathbb{R}} \left\{ |\phi'_k(0)|, \frac{|\phi'_k(0)|^2}{L_{k,0} + \eta \min\{0, \phi''_k(0)\}} \right\}
\]

\[
= \lim_{k \to \infty} \min_{\kappa \in \mathbb{R}} \left\{ |g_k^T(p_k + d_k)|, \frac{|g_k^T(p_k + d_k)|^2}{L||p_k + d_k||_2^2 + \eta \min\{0, (p_k + d_k)^TH_k(p_k + d_k)\}} \right\}
\]

\[
= \lim_{k \to \infty} \min_{\kappa \in \mathbb{R}} \left\{ |g_k^T p_k|, \frac{|g_k^T p_k|^2}{L||p_k + d_k||_2^2 + \eta \min\{0, \lambda_k||p_k + d_k||_2^2\}} \right\}
\]

\[
= \lim_{k \to \infty} \min_{\kappa \in \mathbb{R}} \left\{ \left(\gamma_p \frac{||g_k||_2^2}{||p_k||_2^2}\right), \frac{\gamma_p^2 ||g_k||_2^2||p_k||_2^2}{L||p_k + d_k||_2^2 + \eta \min\{0, \lambda_k||p_k + d_k||_2^2\}} \right\}
\]

\[
= \lim_{k \to \infty} \min_{\kappa \in \mathbb{R}} \left\{ \left(\frac{\gamma_p}{\kappa_p}\right) ||g_k||_2^2, \frac{\left(\frac{\gamma_p}{\kappa_p}\right)^2 ||g_k||_2^4}{L||p_k + d_k||_2^2 + \eta \min\{0, \lambda_k||p_k + d_k||_2^2\}} \right\}.
\]

If we assume that \(\{g_k\}\) is uniformly bounded and that \(\{\lambda_k\}\) is uniformly bounded from below, then it follows from (17) and (19) that \(\{||p_k||_2^2\}\) and \(\{||d_k||_2^2\}\) are uniformly bounded, which in turn imply that there exists a number \(M_1 \in (0, \infty)\) such that

\[
L||p_k + d_k||_2^2 + \eta \min\{0, \lambda_k||p_k + d_k||_2^2\} \leq M_1 \quad \text{for all } k.
\]

Combining this with the above limit shows that

\[
\lim_{k \to \infty} ||g_k||_2 = 0.
\]
We know from the previous limit and (17) that
\[
\lim_{k \to \infty} \|g_k\|_2 = 0 \quad \text{and} \quad \lim_{k \to \infty} \|p_k\|_2 = 0
\] (44)
which is the first main result that we wish to prove.

Next, it follows from (41) and (43) that
\[
0 = \lim_{k \to \infty} \min \left\{ \left| \min\{0, \phi''_k(0)\} \right|, \frac{\left| \min\{0, \phi''_k(0)\} \right|^3}{(\sigma_k, 0)^2} \right\}
\]
\[
= \lim_{k \to \infty} \min \left\{ \left| \min\{0, (p_k + d_k)^T H_k(p_k + d_k)\} \right|, \frac{\left| \min\{0, (p_k + d_k)^T H_k(p_k + d_k)\} \right|^3}{\sigma^2 \|p_k + d_k\|_2^6} \right\}
\]
Using, again, the uniform boundedness of \( \{p_k\} \) and \( \{d_k\} \) we know that there exists a constant \( M_2 \in (0, \infty) \) such that
\[
\sigma^2 \|p_k + d_k\|_2^6 \leq M_2 \quad \text{for all } k.
\]
Combining this with the above limit gives
\[
0 = \lim_{k \to \infty} \min \left\{ \left| \min\{0, (p_k + d_k)^T H_k(p_k + d_k)\} \right|, \frac{\left| \min\{0, (p_k + d_k)^T H_k(p_k + d_k)\} \right|^3}{M_2} \right\}
\]
We now aim to prove that
\[
\lim_{k \to \infty} \|d_k\|_2 = 0. \tag{45}
\]
For a proof by contradiction, suppose that there exists a scalar \( \epsilon > 0 \) such that the set
\[
\mathcal{D} := \{ k : \|d_k\|_2 \geq \epsilon \}
\]
is infinite. Observe from (19) and the definition of \( \mathcal{D} \) that
\[
\epsilon \leq \|d_k\|_2 \leq -\kappa_d \lambda_k \quad \text{for all } k \in \mathcal{D}
\]
which may be combined with (18) and the definition of \( \mathcal{D} \) to conclude that
\[
d_k^T H_k d_k \leq \gamma_d \lambda_k \|d_k\|_2^2 \leq \gamma_d \lambda_k \epsilon^2 \leq -\left( \frac{\gamma_d}{\kappa_d} \right) \epsilon^3 < 0 \quad \text{for all } k \in \mathcal{D}.
\]
(46)

If we assume that \( \{H_k\} \) is uniformly bounded, then the previous inequality may be combined with (44) to show that
\[
(p_k + d_k)^T H_k (p_k + d_k) \leq \frac{1}{2} d_k^T H_k d_k < 0 \quad \text{for all sufficiently large } k \in \mathcal{D}.
\]
Combining this with the limit on the previous slide and (46) shows that
\[
0 = \lim \min_{k \in \mathcal{D}} \left\{ \left| \min\{0, \frac{1}{2} d_k^T H_k d_k\} \right|, \frac{\left| \min\{0, \frac{1}{2} d_k^T H_k d_k\} \right|^3}{M_2} \right\}
= \lim \min_{k \in \mathcal{D}} \left\{ \left| \min\{0, -\frac{1}{2} \left( \frac{\gamma_d}{\kappa_d} \right) \epsilon^3\} \right|, \frac{\left| \min\{0, -\frac{1}{2} \left( \frac{\gamma_d}{\kappa_d} \right) \epsilon^3\} \right|^3}{M_2} \right\}
= \lim \min_{k \in \mathcal{D}} \left\{ \frac{1}{2} \left( \frac{\gamma_d}{\kappa_d} \right) \epsilon^3, \frac{\left[ \frac{1}{2} \left( \frac{\gamma_d}{\kappa_d} \right) \epsilon^3 \right]^3}{M_2} \right\} = \min \left\{ \frac{1}{2} \left( \frac{\gamma_d}{\kappa_d} \right) \epsilon^3, \frac{\left[ \frac{1}{2} \left( \frac{\gamma_d}{\kappa_d} \right) \epsilon^3 \right]^3}{M_2} \right\} > 0
\]
which is a contradiction. Thus, we have proved that (45) holds.
Next, it follows from (45) and (19) that

$$\lim_{k \to \infty} \left| \min\{0, \lambda_k\} \right| = 0$$

which means that

$$\liminf_{k \to \infty} \lambda_k \geq 0$$

thus giving our second desired result.

We have proved the following result.
Theorem 14

Let the following assumptions on the objective function and iterates hold:

- \( f \) is twice continuously differentiable;
- \( \nabla f \) and \( \nabla^2 f \) are both Lipschitz continuous; and
- the sequences \( \{g_k\} \) and \( \{H_k\} \) are uniformly bounded.

If the weighting functions
\[
\omega_{k,p}(\alpha) = \omega_{k,d}(\alpha) = \alpha
\]
are used for all \( k \), then one of the following outcomes must hold for Algorithm 3:

(i) finite termination, i.e., there exists an iterate \( x_k \) such that
\[
g_k = 0 \quad \text{and} \quad \lambda_k \geq 0
\]

(ii) unbounded objective, i.e.,
\[
\lim_{k \to \infty} f(x_k) = -\infty
\]

(iii) second-order optimality in the limit, i.e, the iterates satisfy
\[
\lim_{k \to \infty} \|g_k\|_2 = 0 \quad \text{and} \quad \liminf_{k \to \infty} \lambda_k \geq 0
\]

so that limit points of \( \{x_k\} \) satisfy second-order necessarily optimality conditions.

Proof: The proof is given by our previous discussion.
1 Introduction

2 Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3 Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4 Trust-Region Methods

5 Conclusions and Final Thoughts
Maybe next year . . . but for now you can see [1]!
Outline

1 Introduction

2 Fixed Step Size Algorithms
   - Convergence to first-order solutions
   - Convergence to second-order solutions

3 Line Search Methods
   - Introduction
   - Algorithmic framework
   - Analysis of the framework
   - An algorithmic instance: equal weighting

4 Trust-Region Methods

5 Conclusions and Final Thoughts
Final comments

- Fixed step size approach can ensure limit points satisfy second-order optimality.
- Relatively little research has focused on second-order methods.
  - They have a stronger convergence theory.
  - Extra expense to compute negative curvature direction $d_k$.
  - Little empirical gain (if any) in reducing the number of iterations.
  - More research is needed, in my opinion.

- Is it worth it to computing a direction of negative curvature?
  - Many (most?) optimization experts would say no!
  - Personally, I am not sure, but I agree that current methods do not benefit enough from negative curvature directions to justify their use.
  - I believe we need better methods. How? This is ongoing research!

- Fixed step size can be quite powerful if the global Lipschitz constant is similar in size to the local Lipschitz constants encountered over the sequence $\{x_k\}$.

- Curvilinear searches are rarely used in practice. Empirical performance gains are minimal or altogether missing.
