1 Introduction

In this paper, we are interested in developing and analyzing the convergence of algorithms that solve the optimization problem

\[
\begin{align*}
\min_{x} & \quad f(x) \\
\text{s.t.} & \quad x \in \mathcal{X} \subset \mathbb{R}^l
\end{align*}
\]  

(P)

where the objective function \( f \) is sufficiently smooth. We consider iterative search algorithms, i.e., algorithms that start with a guess \( x_0 \in \mathcal{X} \) of a solution of (P) and successively generate points \( x_1, x_2, \ldots \), such that the sequence \( \{x_n\}_{n \in \mathbb{N}} \) converges to an optimal solution of (P). There are a wide variety of such iterative algorithms in the literature that use information of the function values \( f(x_n) \) and the values \( \nabla f(x_n) \) and \( \nabla^2 f(x_n) \) of the first and second derivatives at each point \( x_n \), in order to generate the next point \( x_{n+1} \).

There are many important optimization problems for which the exact evaluation of the objective function \( f(x) \) and the derivatives \( \nabla f(x) \) and \( \nabla^2 f(x) \) at a given point \( x \) requires a lot of effort; and sometimes it is for practical purposes impossible to evaluate these quantities exactly. An example of such a problem is a stochastic optimization problem, where the objective function \( f \) is of the form

\[
f(x) := \mathbb{E}[F(x, \zeta)]
\]  

(1)
where the expectation is with respect to the distribution of the random variable $\zeta$. Often it is very hard or impossible to compute the expected value $\mathbb{E}[F(x, \zeta)]$ at a particular point $x$ exactly, especially if $\zeta$ is high dimensional. Another example of such a problem is an optimization problem in which $f$ is given by the solution of a differential equation for which a closed-form solution is not available.

It is often the case that for such objective functions it is possible construct an approximation $\hat{f}(\cdot, N)$ of $f$ where the parameter $N$ controls the accuracy of the approximation. Consequently, one approach to such optimization problems is to create a sequence $\{f_n\}_{n \in \mathbb{N}}$ of approximations of $f$ using $\hat{f}(\cdot, N)$. The idea is to obtain approximations $f_n$ that are easier to compute than the original function $f$, and to control the approximation error as needed. Note that the approximations $f_n$ may be random. For example, for the stochastic optimization problem with an objective function as defined in (1), an approximation $\hat{f}(\cdot, N)$ can be obtained by sampling, as follows: Let $\zeta_1, \ldots, \zeta_N$ be a sample of values of the random variable $\zeta$, Then the sample average approximation is given by

$$
\hat{f}(x, N) := \frac{1}{N} \sum_{i=1}^{N} F(x, \zeta_i)
$$

(2)

It follows from strong laws of large numbers that, under some conditions, the approximation error can be controlled by controlling the sample size $N$. We define $f_n$ by picking a sample size $N_n$ and setting $f_n = \hat{f}(\cdot, N_n)$. Values $f_n(x)$ of the sample average approximation are easy to compute if $N_n$ is not too large, it is easy to generate sample observations $\zeta_i$, and it is easy to compute $F(x, \zeta_i)$ for any given $x$ and $\zeta_i$. Another example, for the case in which $f$ is given by the solution of a differential equation, is a sequence $\{f_n\}_{n \in \mathbb{N}}$ of approximations based on a sequence of discretizations.

We see that often we are able to approximate the objective function $f$ with functions $f_n$, and we are able to control the approximation error (with effort). In some of these cases it may be easy to calculate the derivatives of the approximating functions $f_n$ as well, and these may provide good approximations of the derivatives of the objective function $f$. However, it may also happen that $f_n$ is not differentiable at some points, or the derivative values of $f_n$ may be hard to compute at given points. For example, in the stochastic optimization problem, $F$ may not be differentiable with respect to $x$ at all points $x$ for all $\zeta_i$, and thus $f_n$ may not be differentiable at all $x$. An example where this is the case is $F(x, \zeta) := |x - \zeta|$. In this case, $F$ is not differentiable with respect to $x$ at $x = \zeta$, and $f_n$ is not differentiable at the points $x = \zeta_1, \ldots, \zeta_N$. However, if $\zeta$ has a continuous
distribution, then \( f \) is continuously differentiable with derivative \( f'(x) = 2P[\zeta \leq x] - 1 \). Also, it may be relatively easy to implement a simulation to compute the sample average approximation \( f_n(x) \) in (2), but even if \( f_n \) is differentiable at \( x \), there may be several practical reasons why one does not want to implement the code to compute the derivatives of \( f_n \) at \( x \). It may require much more work to implement a simulation, and verify the correctness of the implementation, to also compute the derivatives of \( f_n \), and the chance of implementation errors may be much larger as well. Also, in several applications, the sampled objective values \( F(x, \zeta) \) at given points \((x, \zeta)\) are provided by protected code, which is treated as a black box by the optimizer. Thus, taking into account practical considerations, even if the approximations \( f_n \) are differentiable, one may prefer to implement a relatively simple simulation to compute \( f_n(x) \) but not the derivatives of \( f_n \), and to use an algorithm that does not require such derivatives, even if the algorithm is less efficient than an algorithm that requires derivatives.

In this paper we consider iterative algorithms that solve the optimization problem (P), using sequences of approximations \( \{f_n\}_{n \in \mathbb{N}}, \{\hat{\nabla} f_n\}_{n \in \mathbb{N}}, \) and possibly \( \{\hat{\nabla}^2 f_n\}_{n \in \mathbb{N}} \), of the objective function \( f \), its gradient \( \nabla f \), and possibly its Hessian \( \nabla^2 f \), respectively. In Section 3, we look at an objective function \( f \) of the form (1), and show the convergence properties of the sample average approximation \( \hat{f}(\cdot,N) \) as \( N \to \infty \). In Section 4, we show how the sequences \( \{f_n\}_{n \in \mathbb{N}}, \{\hat{\nabla} f_n\}_{n \in \mathbb{N}} \) and \( \{\hat{\nabla}^2 f_n\}_{n \in \mathbb{N}} \) approximating \( f \), \( \nabla f \) and \( \nabla^2 f \) respectively, can be generated using local linear regression and \( \hat{f}(\cdot,N) \). In Sections 5 through 8, we define and prove the convergence of the various iterative optimization algorithms that use these approximating function and gradient sequences.

### 2 Notation and Some Mathematical Background

In this section, we define the notation that will be used in the rest of this document and we also state and prove some general results that will be used later.

Let \( \mathbb{N} \) denote the set of natural numbers and let \( \mathbb{Z} \) denote the set of integers. In this paper, we deal with finite dimensional real vector spaces. Let \( \mathbb{R}^l \) denote a real vector space of dimension \( l \in \mathbb{N} \). Any norm on a vector space is denoted with \( \| \cdot \| \). We use the following notation for various
widely used norms: For $x \in \mathbb{R}^l$, and $p \in [1, \infty)$, let
\[
\|x\|_p := \left( \sum_{i=1}^l |x_i|^p \right)^{1/p}
\]
and let
\[
\|x\|_\infty := \max_{i \in \{1, \ldots, l\}} |x_i|
\]
It is well known that all norms on any finite dimensional vector space are equivalent. That is, for any two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^l$, there exist constants $0 < c_{ab}^1 < c_{ab}^2 < \infty$ such that for any $x \in \mathbb{R}^l$,
\[
c_{ab}^1 \|x\|_a \leq \|x\|_b \leq c_{ab}^2 \|x\|_a
\]
In particular, the following inequalities relating various norms are well-known. For any $x \in \mathbb{R}^l$,
\[
\|x\|_2 \leq \|x\|_1 \leq \sqrt{l} \|x\|_2 \tag{3}
\]
\[
\|x\|_\infty \leq \|x\|_2 \leq \sqrt{l} \|x\|_\infty \tag{4}
\]
For any $x \in \mathbb{R}^l$ and $\delta > 0$, let
\[
B(x, \delta) := \left\{ y \in \mathbb{R}^l : \|y - x\|_2 \leq \delta \right\}
\]
denote the closed ball of radius $\delta$ centered at $x$. The following is a useful result regarding the topological structure of $\mathbb{R}^l$.

**Lemma 2.1.** Consider an open set $\mathcal{E} \subset \mathbb{R}^l$ and a compact set $\mathcal{D} \subset \mathcal{E}$. Then there exists $\delta_\mathcal{D} > 0$ such that for all $x \in \mathcal{D}$, $B(x, \delta_\mathcal{D}) \subset \mathcal{E}$. Further, let
\[
\mathcal{D}^* := \bigcup_{x \in \mathcal{D}} B(x, \delta_\mathcal{D})
\]
Then, the set $\mathcal{D}^*$ is compact and satisfies $\mathcal{D} \subset \mathcal{D}^* \subset \mathcal{E}$.

**Proof.** We show this result by contradiction. Suppose that the first assertion of Lemma 2.1 is not true. Then there exists a sequence $\{(x_n, \delta_n)\}_{n \in \mathbb{N}}$ with each $x_n \in \mathcal{D}$ and $\delta_n \downarrow 0$ as $n \to \infty$ such that $B(x_n, \delta_n) \notin \mathcal{E}$. Since $\mathcal{D}$ is a compact set, there is a subsequence $\{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N}$ such that $x_{n_k} \to \tilde{x} \in \mathcal{D} \subset \mathcal{E}$ as $k \to \infty$. 

4
Now, since $E$ is an open set, there exists $\tilde{\delta} > 0$ such that $B(\tilde{x}, \tilde{\delta}) \subset E$. Also, since $x_{n_k} \to \tilde{x}$ as $k \to \infty$, there exists $N_1 \in \mathbb{N}$ such that for all $k > N_1$, $x_{n_k} \in B(\tilde{x}, \tilde{\delta}/2)$. This implies that for all $k > N_1$,

$$B(x_{n_k}, \tilde{\delta}/2) \subset B(\tilde{x}, \tilde{\delta}) \subset E$$

Finally, let $N_2 \in \mathbb{N}$ be such that for all $k > N_2$, $\delta_{n_k} < \tilde{\delta}/2$. Then, for all $k > \max\{N_1, N_2\}$,

$$B(x_{n_k}, \delta_{n_k}) \subset B(x_{n_k}, \tilde{\delta}/2) \subset B(\tilde{x}, \tilde{\delta}) \subset E$$

This contradicts the assumption that $B(x_n, \delta_n) \not\subset E$ for each $n \in \mathbb{N}$. Therefore, there exists $\delta_D > 0$ such that $B(x, \delta_D) \subset E$ for all $x \in D$.

Consider the set $D^* := \bigcup_{x \in D} B(x, \delta_D)$. Clearly since $x \in B(x, \delta_D)$ for each $x \in D$, it holds that $D \subset D^*$. Also, since $B(x, \delta_D) \subset E$ for each $x \in D$, it holds that $D^* \subset E$. It remains to show that $D^*$ is a compact set.

We first show that $D^*$ is a bounded set. Consider any $y \in D^*$. It follows from the definition of $D^*$ that there exists $x \in D$ such that $y \in B(x, \delta_D)$. Thus,

$$\|y\|_2 \leq \|x\|_2 + \|y - x\|_2 \leq \sup_{x \in D} \|x\|_2 + \delta_D$$

Thus, since $D$ is compact and hence bounded, it follows that $D^*$ is a bounded set.

Next, we show that $D^*$ is closed. Consider any sequence $\{y_n\}_{n \in \mathbb{N}} \subset D^*$ converging to some $\tilde{y} \in \mathbb{R}^l$. For each $n \in \mathbb{N}$, there exists $x_n \in D$ such that $y_n \in B(x_n, \delta_D)$. Then, since $D$ is compact, there is a subsequence $\{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N}$ and $\tilde{x} \in D$ such that $x_{n_k} \to \tilde{x}$ as $k \to \infty$. Also, $y_{n_k} \to \tilde{y}$ as $k \to \infty$. But for each $k \in \mathbb{N}$, $\|y_{n_k} - x_{n_k}\|_2 \leq \delta_D$. Hence

$$\|\tilde{y} - \tilde{x}\|_2 = \lim_{k \to \infty} \|y_{n_k} - x_{n_k}\|_2 \leq \delta_D$$

Therefore, $\tilde{y} \in B(\tilde{x}, \delta_D) \subset D^*$. Thus, $D^*$ is a closed set and consequently a compact set.

\begin{lemma}
Consider a sequence $\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \subset \mathbb{R}^l$. Let $\mathcal{A}$ denote the set of accumulation points of $\{x_n\}_{n \in \mathbb{N}}$, and let $\mathcal{U} := \bigcup_{n \in \mathbb{N}}\{x_n\}$ denote the set of points in $\{x_n\}_{n \in \mathbb{N}}$. Then, $\overline{\mathcal{U}} = \mathcal{A} \cup \mathcal{U}$. Consequently, if $\{x_n\}_{n \in \mathbb{N}}$ is bounded and $\mathcal{A} \subset \mathcal{X}$, then $\overline{\mathcal{U}}$ is a compact subset of $\mathcal{X}$.
\end{lemma}

5
Proof. First, we show that \( \text{cl}(\mathcal{U}) = A \cup \mathcal{U} \). Recall that
\[
\text{cl}(\mathcal{U}) := \left\{ x \in \mathbb{R}^l : x = \lim_{k \to \infty} \tilde{x}_k \text{ where } \{\tilde{x}_k\}_{k \in \mathbb{N}} \subset \mathcal{U} \right\}
\]
It is clear that \( \mathcal{U} \subset \text{cl}(\mathcal{U}) \) and \( A \subset \text{cl}(\mathcal{U}) \). Thus \( A \cup \mathcal{U} \subset \text{cl}(\mathcal{U}) \).

Next, consider any \( x \in \text{cl}(\mathcal{U}) \). Thus there is a sequence \( \{\tilde{x}_k\}_{k \in \mathbb{N}} \subset \mathcal{U} \) such that \( x = \lim_{k \to \infty} \tilde{x}_k \).

Then, for each \( k \in \mathbb{N} \), there is \( n_k \in \mathbb{N} \) such that \( \tilde{x}_k = x_{n_k} \). Now, two cases arise with respect to the properties of \( \{n_k\}_{k \in \mathbb{N}} \).

Case (a): Suppose \( \{n_k\}_{k \in \mathbb{N}} \) is bounded, that is, there exists \( M \in \mathbb{N} \) such that \( n_k < M \) for each \( k \in \mathbb{N} \). Then there is \( n^* < M \) and \( N \in \mathbb{N} \) such that \( \tilde{x}_k = x_{n^*} \) for all \( k > N \). Consequently \( x = x_{n^*} \in \mathcal{U} \).

Case (b): Suppose \( \{n_k\}_{k \in \mathbb{N}} \) is unbounded, that is, there is a subsequence \( \{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}} \) such that \( \lim_{k \to \infty} m_k = \infty \). Note that \( \lim_{k \to \infty} x_{m_k} = \lim_{k \to \infty} x_{n_k} = \lim_{k \to \infty} \tilde{x}_k = x \). Then it follows from the definition of \( A \) that \( x \in A \).

Thus, \( x \in A \cup \mathcal{U} \); hence, \( \text{cl}(\mathcal{U}) \subset A \cup \mathcal{U} \). Consequently \( \text{cl}(\mathcal{U}) = A \cup \mathcal{U} \).

Suppose that \( \{x_n\}_{n \in \mathbb{N}} \) is bounded and \( A \subset X \). Since \( \{x_n\}_{n \in \mathbb{N}} \) is bounded, it follows that \( A \) is bounded as well. Thus \( A \cup \mathcal{U} \) is a bounded subset of \( X \). Also, it was shown above that \( A \cup \mathcal{U} = \text{cl}(\mathcal{U}) \), and thus \( \text{cl}(\mathcal{U}) \) is a closed and bounded subset of \( X \). Therefore, \( \text{cl}(\mathcal{U}) \) is a compact subset of \( X \). \( \square \)

For any \( l, m \in \mathbb{N} \), let \( \mathbb{R}^{m \times l} \) denote the space of \( m \times l \) real matrices. We use the following notation for widely used matrix norms. For any \( A = (a_{ij}) \in \mathbb{R}^{m \times l} \) and \( p \in [1, \infty) \), let
\[
\|A\|_p := \sup_{\{x \in \mathbb{R}^l : \|x\|_p = 1\}} \|Ax\|_p
\]
Also, let
\[
\|A\|_\infty := \max_{i \in \{1, \ldots, m\}} \sum_{j=1}^l |a_{ij}|
\]
\[
\|A\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^l |a_{ij}|^2}
\]
The following inequalities involving various matrix norms are well known. For any $A \in \mathbb{R}^{m \times l}$,

$$
\|A\|_2 \leq \|A\|_p \leq \sqrt{l} \|A\|_2
$$

$$
\frac{1}{\sqrt{l}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{m} \|A\|_\infty
$$

$$
\frac{1}{\sqrt{m}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{l} \|A\|_1
$$

Further, for any $A \in \mathbb{R}^{m \times l}$, $B \in \mathbb{R}^{l \times n}$ and $p \in [0, \infty)$, we know that

$$
\|AB\|_p \leq \|A\|_p \|B\|_p
$$

In this paper, we will often deal with symmetric matrices. Accordingly, we let $S^{l \times l}$ denote the set of $l \times l$ symmetric real matrices. For any $S = (s_{ij}) \in S^{l \times l}$, let $\lambda_1(S), \ldots, \lambda_l(S)$ denote the eigenvalues of $S$, and let $\lambda_{\text{max}}(S) := \max\{\lambda_1(S), \ldots, \lambda_l(S)\}$ and $\lambda_{\text{min}}(S) := \min\{\lambda_1(S), \ldots, \lambda_l(S)\}$ denote the maximum and minimum eigenvalues of $S$ respectively. It is well known that

$$
\lambda_{\text{max}}(S) := \max\{x \in \mathbb{R}^l : \|x\|_2 = 1\} x^T S x
$$

$$
\lambda_{\text{min}}(S) := \min\{x \in \mathbb{R}^l : \|x\|_2 = 1\} x^T S x
$$

Consequently, for any $x \in \mathbb{R}^l$,

$$
\lambda_{\text{min}}(S) \|x\|_2^2 \leq x^T S x \leq \lambda_{\text{max}}(S) \|x\|_2^2
$$

Further

$$
\|S\|_2 := \max\{x \in \mathbb{R}^l : \|x\|_2 = 1\} \|Sx\|_2 = \max\{|\lambda_{\text{max}}(S)|, |\lambda_{\text{min}}(S)|\}
$$

It follows that for any $x \in \mathbb{R}^l$,

$$
|x^T S x| \leq \max\{|\lambda_{\text{max}}(S)|, |\lambda_{\text{min}}(S)|\} \|x\|_2^2 = \|S\|_2 \|x\|_2^2
$$

The trace of any square matrix $S$ is defined as follows:

$$
\text{trace}(S) := \sum_{j=1}^l s_{jj} = \sum_{j=1}^l \lambda_j(S)
$$

Let $S_{+}^{l \times l}$ and $S_{++}^{l \times l}$ denote the set of positive semidefinite and positive definite $l \times l$ symmetric matrices respectively. If $S \in S_{+}^{l \times l}$, then $\lambda_j(S) \geq 0$ for all $j \in \{1, \ldots, l\}$. Therefore,

$$
\|S\|_2 \leq \text{trace}(S) \leq l \|S\|_2 \quad \text{for all } S \in S_{+}^{l \times l}
$$
We will deal only with real valued functions whose domains are subsets of $\mathbb{R}^l$ (for some $l \in \mathbb{N}$). We denote such functions $f, g, h$ and so on. Let $\mathcal{E} \subseteq \mathbb{R}^l$ be an open set. Then $\mathcal{C}_0(\mathcal{E})$ denotes the space of continuous real-valued functions on $\mathcal{E}$. Similarly, $\mathcal{C}_n(\mathcal{E})$ denotes the space of $n$-times continuously differentiable real-valued functions on $\mathcal{E}$. Also, $\mathcal{W}_0(\mathcal{E})$ denotes the space of locally Lipschitz continuous functions on $\mathcal{E}$ and $\mathcal{W}_n(\mathcal{E})$ denotes the space of $n$-times locally Lipschitz continuously differentiable functions on $\mathcal{E}$. Then,

$$
\mathcal{C}_0(\mathcal{E}) \supset \mathcal{W}_0(\mathcal{E}) \supset \mathcal{C}_1(\mathcal{E}) \supset \mathcal{W}_1(\mathcal{E}) \supset \mathcal{C}_2(\mathcal{E}) \ldots
$$

Next, we present two results regarding functions in $\mathcal{C}_1(\mathcal{E})$ and $\mathcal{C}_2(\mathcal{E})$.

**Lemma 2.3.** Consider an open set $\mathcal{E} \subseteq \mathbb{R}^l$ and a function $f \in \mathcal{C}_1(\mathcal{E})$. Let $\mathcal{D}$ be a compact subset of $\mathcal{E}$. Then there exists a constant $K_{1f} < \infty$ such that

$$
\sup_{\{x,y \in \mathcal{D} : x \neq y\}} \frac{|f(y) - f(x) - \nabla f(x)^T(y - x)|}{\|y - x\|_2} < K_{1f}
$$

**Proof.** By contradiction. Suppose that the assertion in Lemma 2.3 does not hold. Then there exists a sequence $\{(x_n, y_n)\}_{n \in \mathbb{N}} \subset \mathcal{D}^2$ with $x_n \neq y_n$ for each $n \in \mathbb{N}$, such that

$$
\lim_{n \to \infty} \frac{|f(y_n) - f(x_n) - \nabla f(x_n)^T(y_n - x_n)|}{\|y_n - x_n\|_2} = \infty \tag{8}
$$

Since $\mathcal{D}$ is compact, there is a $(\tilde{x}, \tilde{y}) \in \mathcal{D}^2$ and a subsequence of $\{(x_n, y_n)\}_{n \in \mathbb{N}}$, also denoted with $\{(x_n, y_n)\}_{n \in \mathbb{N}}$, such that $x_n \to \tilde{x}$ and $y_n \to \tilde{y}$ as $n \to \infty$. Next we consider two cases.

**Case a:** $\tilde{x} \neq \tilde{y}$: Then, since $f$ and $\nabla f$ are continuous on $\mathcal{D}$, it holds that

$$
\lim_{n \to \infty} \frac{|f(y_n) - f(x_n) - \nabla f(x_n)^T(y_n - x_n)|}{\|y_n - x_n\|_2} = \frac{|f(\tilde{y}) - f(\tilde{x}) - \nabla f(\tilde{x})^T(\tilde{y} - \tilde{x})|}{\|\tilde{y} - \tilde{x}\|_2} < \infty
$$

which contradicts (8).

**Case b:** $\tilde{x} = \tilde{y}$: Then $\|y_n - x_n\|_2 \to 0$ as $n \to \infty$. Since $\mathcal{D} \subset \mathcal{E}$ is compact, it follows from Lemma 2.1 that there exists $\delta_D > 0$ such that $\mathcal{B}(x, \delta_D) \subset \mathcal{E}$ for all $x \in \mathcal{D}$. In particular, $\mathcal{B}(x_n, \delta_D) \subset \mathcal{E}$ for all $n \in \mathbb{N}$. Since $\|y_n - x_n\|_2 \to 0$ as $n \to \infty$, there exists $N \in \mathbb{N}$ such that $y_n \in \mathcal{B}(x_n, \delta_D)$ for all $n > N$. Because $\mathcal{B}(x_n, \delta_D)$ is convex, the line segment $[x_n, y_n]$ is contained in $\mathcal{B}(x_n, \delta_D) \subset \mathcal{E}$ for all $n > N$. Since $f$ is continuously differentiable on $\mathcal{E}$, it follows from the mean value theorem that for each $n > N$,

$$
f(y_n) - f(x_n) = \nabla f(x_n + t_n(y_n - x_n))^T(y_n - x_n) \quad \text{for some} \quad t_n \in [0, 1]
$$
Therefore, for all $n > N$, 
\[
\frac{|f(y_n) - f(x_n) - \nabla f(x_n)^T(y_n - x_n)|}{\|y_n - x_n\|_2} = \frac{|(\nabla f(x_n + t_n(y_n - x_n)) - \nabla f(x_n))^T(y_n - x_n)|}{\|y_n - x_n\|_2} \leq \|\nabla f(x_n + t_n(y_n - x_n)) - \nabla f(x_n)\|_2
\]

The inequality above follows from the Cauchy-Schwarz inequality and the assumption that $x_n \neq y_n$.

Recall that $x_n \to \bar{x}$ and $y_n \to \bar{x}$ as $n \to \infty$, and thus $x_n + t_n(y_n - x_n) \to \bar{x}$ as $n \to \infty$. Therefore, since $f$ is continuously differentiable on $E$, 
\[
\lim_{n \to \infty} \|\nabla f(x_n + t_n(y_n - x_n)) - \nabla f(x_n)\|_2 \leq \lim_{n \to \infty} \|\nabla f(x_n + t_n(y_n - x_n)) - \nabla f(\bar{x})\|_2 + \lim_{n \to \infty} \|\nabla f(\bar{x}) - \nabla f(x_n)\|_2 = 0
\]

Therefore, 
\[
\lim_{n \to \infty} \left| \frac{f(y_n) - f(x_n) - \nabla f(x_n)^T(y_n - x_n)}{\|y_n - x_n\|_2} \right| = 0
\]

This contradicts (8).

\[\Box\]

**Lemma 2.4.** Consider an open set $E \subset \mathbb{R}^l$ and a function $f \in C^2(E)$. Let $D$ be a compact subset of $E$. Then there exists a constant $K_{2f} < \infty$ such that 
\[
\sup_{\{x, y \in D : x \neq y\}} \left| \frac{f(y) - f(x) - \nabla f(x)^T(y - x) - \frac{1}{2}(y - x)^T\nabla^2 f(x)(y - x)}{\|y - x\|_2^2} \right| < K_{2f}
\]

**Proof.** By contradiction. Suppose that the assertion in Lemma 2.4 does not hold. Then there exists a sequence $\{(x_n, y_n)\}_{n \in \mathbb{N}} \subset D^2$ with $x_n \neq y_n$ for each $n \in \mathbb{N}$, such that 
\[
\lim_{n \to \infty} \left| \frac{f(y_n) - f(x_n) - \nabla f(x_n)^T(y_n - x_n) - \frac{1}{2}(y_n - x_n)^T\nabla^2 f(x_n)(y_n - x_n)}{\|y_n - x_n\|_2^2} \right| = \infty \quad (9)
\]

Since $D$ is compact, there is a $(\bar{x}, \bar{y}) \in D^2$ and a subsequence of $\{(x_n, y_n)\}_{n \in \mathbb{N}}$, also denoted with $\{(x_n, y_n)\}_{n \in \mathbb{N}}$, such that $x_n \to \bar{x}$ and $y_n \to \bar{y}$ as $n \to \infty$. Next we consider two cases.
Case a: $\tilde x \neq \tilde y$: Then, since $f \in \mathbb{C}_2(\mathcal{E})$,
\[
\lim_{n \to \infty} \frac{|f(y_n) - f(x_n) - \nabla f(x_n)^T (y_n - x_n) - \frac{1}{2} (y_n - x_n)^T \nabla^2 f(x_n) (y_n - x_n)|}{\|y_n - x_n\|_2^2} = \frac{|f(\tilde y) - f(\tilde x) - \nabla f(\tilde x)(\tilde y - \tilde x) - \frac{1}{2} (\tilde y - \tilde x)^T \nabla^2 f(\tilde x)(\tilde y - \tilde x)|}{\|\tilde y - \tilde x\|_2^2} < \infty
\]
This contradicts (9).

Case b: $\tilde x = \tilde y$: Then $\|y_n - x_n\|_2 \to 0$ as $n \to \infty$. Since $\mathcal{D} \subset \mathcal{E}$ is compact, it follows from Lemma 2.1 that there exists $\delta_\mathcal{D} > 0$ such that $\mathcal{B}(x, \delta_\mathcal{D}) \subset \mathcal{E}$ for all $x \in \mathcal{D}$. In particular, $\mathcal{B}(x_n, \delta_\mathcal{D}) \subset \mathcal{E}$ for all $n \in \mathbb{N}$. Since $\|y_n - x_n\|_2 \to 0$ as $n \to \infty$, there exists $N \in \mathbb{N}$ such that $y_n \in \mathcal{B}(x_n, \delta_\mathcal{D})$ for all $n > N$. Because $\mathcal{B}(x_n, \delta_\mathcal{D})$ is convex, the line segment $[x_n, y_n]$ is contained in $\mathcal{B}(x_n, \delta_\mathcal{D}) \subset \mathcal{E}$ for all $n > N$. Since $f \in \mathbb{C}_2(\mathcal{E})$, it follows from the mean value theorem that for each $n > N$, there exists $t_n \in [0, 1]$ such that
\[
f(y_n) - f(x_n) = \nabla f(x_n)^T (y_n - x_n) + \frac{1}{2} (y_n - x_n)^T \nabla^2 f(x_n + t_n(y_n - x_n))(y_n - x_n)
\]
Therefore, for all $n > N$,
\[
\frac{|f(y_n) - f(x_n) - \nabla f(x_n)^T (y_n - x_n) - \frac{1}{2} (y_n - x_n)^T \nabla^2 f(x_n)(y_n - x_n)|}{\|y_n - x_n\|_2^2} = \frac{1}{2} \left| (y_n - x_n)^T (\nabla^2 f(x_n + t_n(y_n - x_n)) - \nabla^2 f(x_n))(y_n - x_n) \right| \leq \frac{1}{2} \|\nabla^2 f(x_n + t_n(y_n - x_n)) - \nabla^2 f(x_n)\|_2
\]
The inequality above follows from the Cauchy-Schwarz inequality, the definition of matrix norm, and the assumption that $x_n \neq y_n$.

Recall that $x_n \to \tilde x$ and $y_n \to \tilde x$ as $n \to \infty$, and thus $x_n + t_n(y_n - x_n) \to \tilde x$ as $n \to \infty$. Therefore, since $f \in \mathbb{C}_2(\mathcal{E})$,
\[
\lim_{n \to \infty} \|\nabla^2 f(x_n + t_n(y_n - x_n)) - \nabla^2 f(x_n)\|_2 \leq \lim_{n \to \infty} \|\nabla^2 f(x_n + t_n(y_n - x_n)) - \nabla^2 f(\tilde x)\|_2 \leq \lim_{n \to \infty} \|\nabla^2 f(\tilde x) - \nabla^2 f(x_n)\|_2 = 0
\]

Therefore,
\[
\lim_{n \to \infty} \frac{|f(y_n) - f(x_n) - \nabla f(x_n)^T (y_n - x_n) - \frac{1}{2} (y_n - x_n)^T \nabla^2 f(x_n)(y_n - x_n)|}{\|y_n - x_n\|_2^2} = 0
\]
This contradicts (9).
For any compact set $D \subset \mathcal{E}$, $\mathcal{W}_0(D)$ and $\mathcal{W}_1(D)$ are Lipschitz spaces (not merely locally Lipschitz). We can define norms on these spaces as follows. For any $f \in \mathcal{W}_0(D)$, let

$$\|f\|_{\mathcal{W}_0(D)} := \sup_{x \in D} |f(x)| + \sup_{\{x, y \in D : x \neq y\}} \frac{|f(y) - f(x)|}{\|y - x\|_2}.$$  

It follows from Rademacher’s Theorem that for any $f \in \mathcal{W}_0(D)$, there exists a set $\mathcal{N}(f) \subset D$ of Lebesgue measure zero, such that $\nabla f(x)$ exists for all $x \in D \setminus \mathcal{N}(f)$. Further, it also holds that

$$\sup_{\{x, y \in D : x \neq y\}} \frac{|f(y) - f(x)|}{\|y - x\|_2} = \sup_{x \in D \setminus \mathcal{N}(f)} \|\nabla f(x)\|_2$$

Similarly, for any $f \in \mathcal{W}_1(D)$, let

$$\|f\|_{\mathcal{W}_1(D)} := \sup_{x \in D} |f(x)| + \sup_{x \in D} \|\nabla f(x)\|_2 + \sup_{\{x, y \in D : x \neq y\}} \frac{\|\nabla f(y) - \nabla f(x)\|_2}{\|y - x\|_2}$$

Next we establish a useful result regarding the convergence of sequences in $\mathcal{W}_1(D)$.

**Lemma 2.5.** Let $\mathcal{E} \subset \mathbb{R}^d$ be open. Consider a sequence $\{f_n : \mathcal{E} \mapsto \mathbb{R}\}_{n \in \mathbb{N}} \subset \mathcal{W}_1(\mathcal{E})$ and $f \in \mathcal{W}_1(\mathcal{E})$.

If

$$\lim_{n \to \infty} \|f_n - f\|_{\mathcal{W}_1(D)} = 0$$

for any compact set $D \subset \mathcal{E}$, then

$$\lim_{n \to \infty} \sup_{\{x, y \in D : x \neq y\}} \left| \frac{((f_n(y)) - f(y)) - (f_n(x) - f(x)) - (\nabla f_n(x) - \nabla f(x))^T (y-x)}{\|y - x\|_2^2} \right| = 0$$

**Proof.** By contradiction. Suppose that the assertion in Lemma 2.5 does not hold. Then there exists a compact set $D \subset \mathcal{E}$, $\varepsilon > 0$, and subsequences $\mathbb{N}' \subset \mathbb{N}$ and $\{(x_n, y_n)\}_{n \in \mathbb{N}'} \subset D^2$ with $x_n \neq y_n$ for each $n \in \mathbb{N}'$, such that

$$\left| \frac{((f_n(y_n) - f(y_n)) - (f_n(x_n) - f(x_n)) - (\nabla f_n(x_n) - \nabla f(x_n))^T (y_n-x_n))}{\|y_n-x_n\|_2^2} \right| > \varepsilon$$

for all $n \in \mathbb{N}'$. Since $\mathcal{D}$ is compact, there is a $(\tilde{x}, \tilde{y}) \in D^2$ and further subsequences $\mathbb{N}'' \subset \mathbb{N}'$ and $\{(x_n, y_n)\}_{n \in \mathbb{N}''}$ such that $x_n \to \tilde{x}$ and $y_n \to \tilde{y}$ as $n \to \infty$. Next we consider two cases.

**Case a:** $\tilde{x} \neq \tilde{y}$: Then, since $\|f_n - f\|_{\mathcal{W}_1(D)} \to 0$ as $n \to \infty$,

$$f_n(y_n) - f(y_n) \to 0, \ f_n(x_n) - f(x_n) \to 0, \ \text{and} \ \nabla f_n(x_n) - \nabla f(x_n) \to 0$$
Therefore,
\[
\lim_{n \to \infty} \left| \frac{(f_n(y_n) - f(y_n)) - (f_n(x_n) - f(x_n)) - (\nabla f_n(x_n) - \nabla f(x_n))^T(y_n - x_n)}{\|y_n - x_n\|_2^2} \right|
\]
\[
= \frac{1}{\|\tilde{y} - \tilde{x}\|_2^2} \lim_{n \to \infty} \left| (f_n(y_n) - f(y_n)) - (f_n(x_n) - f(x_n)) - (\nabla f_n(x_n) - \nabla f(x_n))^T(\tilde{y} - \tilde{x}) \right| = 0
\]

This contradicts (10).

**Case b:** \(\tilde{x} = \tilde{y}\): Then \(\|y_n - x_n\|_2 \to 0\) as \(n \to \infty\). Since \(D \subset E\) is compact, it follows from Lemma 2.1 that there exists \(\delta_D > 0\) such that \(B(x, \delta_D) \subset E\) for all \(x \in D\), and \(D^* := \bigcup_{x \in D} B(x, \delta_D)\) is compact. In particular, \(B(x_n, \delta_D) \subset E\) for all \(n \in \mathbb{N}\). Since \(\|y_n - x_n\|_2 \to 0\) as \(n \to \infty\), there exists \(N \in \mathbb{N}\) such that \(y_n \in B(x_n, \delta_D)\) for all \(n \in \mathbb{N}', n > N\). Because \(B(x_n, \delta_D)\) is convex, the line segment \([x_n, y_n]\) is contained in \(B(x_n, \delta_D) \subset D^*\) for all \(n \in \mathbb{N}', n > N\).

Let \(g_n \in \mathcal{W}_1(D^*)\) be given by \(g_n(x) = f_n(x) - f(x)\). Then, from the assumption that \(\|f_n - f\|_{\mathcal{W}_1(D^*)} \to 0\) as \(n \to \infty\), it follows that \(\|g_n\|_{\mathcal{W}_1(D^*)} \to 0\) as \(n \to \infty\). Consider any \([x, y] \subset D^*\) such that \(x \neq y\). Then, since \(g_n \in \mathcal{W}_1(D^*) \subset C_1(D^*)\),

\[
g_n(y) - g_n(x) = \int_0^1 \nabla g_n(x + t(y - x))^T(y - x) \, dt
\]

Thus, using the triangle inequality and the Cauchy-Schwarz inequality,

\[
\frac{|g_n(y) - g_n(x) - \nabla g_n(x)^T(y - x)|}{\|y - x\|_2^2} = \frac{\left| \int_0^1 (\nabla g_n(x + t(y - x)) - \nabla g_n(x))^T(y - x) \, dt \right|}{\|y - x\|_2^2}
\]
\[
\leq \int_0^1 \left| (\nabla g_n(x + t(y - x)) - \nabla g_n(x))^T(y - x) \right| \, dt
\]
\[
\leq \int_0^1 \|\nabla g_n(x + t(y - x)) - \nabla g_n(x)\|_2 \|y - x\|_2 \, dt
\]
\[
= \int_0^1 \|\nabla g_n(x + t(y - x)) - \nabla g_n(x)\|_2 \|t(y - x)\|_2 \, dt
\]
\[
\leq \int_0^1 \sup_{\{x^1, x^2 \in D^*: x^1 \neq x^2\}} \left\| \nabla g_n(x^2) - \nabla g_n(x^1) \right\|_2 \|t(y - x)\|_2 \, dt
\]
\[
= \frac{1}{2} \sup_{\{x^1, x^2 \in D^*: x^1 \neq x^2\}} \left\| \nabla g_n(x^2) - \nabla g_n(x^1) \right\|_2 \|x^2 - x^1\|_2
\]

Therefore,

\[
\sup_{\{x, y \subset D^*: x \neq y\}} \left| \frac{g_n(y) - g_n(x) - \nabla g_n(x)^T(y - x)}{\|y - x\|_2^2} \right| \leq \frac{1}{2} \sup_{\{x^1, x^2 \in D^*: x^1 \neq x^2\}} \left\| \nabla g_n(x^2) - \nabla g_n(x^1) \right\|_2 \|x^2 - x^1\|_2
\]
Since $\|g_n\|_{W_1(D^*)} \to 0$ as $n \to \infty$,

$$\lim_{n \to \infty} \sup_{\{[x,y] \subset D^*: x \neq y\}} \left| \frac{g_n(y) - g_n(x) - \nabla g_n(x)^T(y - x)}{\|y - x\|_2^2} \right| \leq \left| \frac{\nabla g_n(x^2) - \nabla g_n(x^1)}{\|x^2 - x^1\|_2^2} \right| = 0$$

In particular,

$$\lim_{n \to \infty} \left| \frac{g_n(y_n) - g_n(x_n) - \nabla g_n(x_n)^T(y_n - x_n)}{\|y_n - x_n\|_2^2} \right| = \lim_{n \to \infty} \left| \frac{(f_n(y_n) - f(y_n)) - (f_n(x_n) - f(x_n)) - (\nabla f_n(x_n) - \nabla f(x_n))^T(y_n - x_n)}{\|y_n - x_n\|_2^2} \right| = 0$$

This contradicts (10).

### 3 Approximating Stochastic Objective Functions

As mentioned in the introduction, in this paper we consider optimization problems in which the objective function values $f(x)$ are difficult or impossible to evaluate exactly. In this section, we consider an important class of such problems where the objective function $f$ is an expected value function as in (1). Specifically, we derive the convergence properties of the sample average function as defined in (2) to $f$.

In order to do this, let $\mathcal{E} \subset \mathbb{R}^l$ be open and $(\Omega, \mathcal{F}, \mathcal{P})$ and $(\Xi, \mathcal{G}, \mathcal{Q})$ be two probability spaces. Let $\{\zeta_n\}_{n \in \mathbb{N}}$ be a sequence of independent and identically distributed random variables on $(\Omega, \mathcal{F}, \mathcal{P})$ taking values in $\Xi$ and let $\mathcal{Q}$ be the probability measure induced on $(\Xi, \mathcal{G})$ by $\zeta_1$. We denote a generic element of $\Xi$ with $\zeta$.

Consider a random real-valued function $F: \mathcal{E} \times \Xi \mapsto \mathbb{R}$ such that for each $x \in \mathcal{E}$, $F(x, \cdot): \Xi \mapsto \mathbb{R}$ is $\mathcal{G}$-measurable. Let the function $f: \mathcal{E} \mapsto \mathbb{R}$ be given by

$$f(x) := \mathbb{E}_\mathcal{Q}[F(x, \zeta)] \quad (11)$$

The sample average function $\hat{f}: \mathcal{E} \times \mathbb{N} \times \Xi \mapsto \mathbb{R}$ corresponding to $f$ as above, can be defined as

$$\hat{f}(x, N)(\omega) := \frac{1}{N} \sum_{j=1}^{N} F(x, \zeta^j(\omega)) \quad (12)$$

Then, the following result regarding the pointwise convergence of $\hat{f}$ to $f$ is well known.
Proposition 3.1. Let \( f: E \rightarrow \mathbb{R} \) be given by (11), and let \( \hat{f}: E \times N \times \Xi \rightarrow \mathbb{R} \) be given by (12). Then, for each \( x \in E \), there exists a set \( N(x) \) with \( P(N(x)) = 0 \) such that for all \( \omega \in \Omega \setminus N(x) \), the following holds. For any \( \epsilon > 0 \) there exists \( M(x, \omega, \epsilon) \) in \( N \) such that

\[
\left| \hat{f}(x, N)(\omega) - f(x) \right| < \epsilon \quad \text{for all} \quad N > M(x, \omega, \epsilon)
\]

Under some stronger conditions on the function \( F \), it is possible to strengthen the convergence of \( \hat{f} \) to \( f \). Accordingly, we begin by noting a few facts regarding the Clarke generalized gradient defined for Lipschitz continuous functions on \( E \).

Consider a function \( g: E \rightarrow \mathbb{R} \) such that

\[
K_g := \sup_{x, y \in E \atop x \neq y} \frac{|g(x) - g(y)|}{\|x - y\|_2} < \infty
\]

The following properties of \( g \) are well known.

\[\text{P 3.1.} \] From Rademacher's Theorem, there exists a set \( U \subset E \) with \( L(U) = 0 \) (where \( L \) denotes the Lebesgue measure on \( \mathbb{R}^l \)) such that \( g \) is Frechet differentiable at all \( x \in E \setminus U \). That is, for any \( x \in E \setminus U \), the vector \( \nabla g(x) = \left( \frac{\partial g}{\partial x_1}(x), \ldots, \frac{\partial g}{\partial x_l}(x) \right)^T \) exists, where \( \frac{\partial g}{\partial x_i}(x) \) denotes the partial derivative of \( g \) with respect to the \( i^{th} \) vector \( e_i \) from the standard basis for \( \mathbb{R}^l \), and satisfies

\[
\lim_{y \to 0} \frac{|g(x + y) - g(x) - y^T \nabla g(x)|}{\|y\|_2} = 0
\]

\[\text{P 3.2.} \] Further, it is known that

\[
\sup_{x \in E \setminus U} \|\nabla g(x)\|_2 = \sup_{x, y \in E \atop x \neq y} \frac{|g(x) - g(y)|}{\|x - y\|_2} = K_g \quad \text{for all} \quad x \in E \setminus U
\]

Consequently, if \( g \in W_0(E) \), then

\[
\|g\|_{W_0(E)} := \sup_{x \in E} |g(x)| + \sup_{x, y \in E \atop x \neq y} \frac{|g(x) - g(y)|}{\|x - y\|_2} = \sup_{x \in E} |g(x)| + \sup_{x \in E \setminus U} \|\nabla g(x)\|_2
\]

\[\text{P 3.3.} \] For any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset E \setminus U \) such that \( x_n \rightarrow x \in E \setminus U \), we have \( \nabla g(x_n) \rightarrow \nabla g(x) \) as \( n \rightarrow \infty \).

Now, for any \( x \in E \), the Clarke generalized gradient \( \partial g(x) \) is a set defined as follows

\[
\partial g(x) := \text{conv} \left\{ v \in \mathbb{R}^l : v = \lim_{k \to \infty} \nabla g(x_k) \quad \text{where} \quad \{x_k\}_{k \in \mathbb{N}} \subset E \setminus U \right\}
\]
P 3.4. For any \( x \in \mathcal{E}, \partial g(x) = \{d(x)\} \) for some \( d(x) \in \mathbb{R}^l \), if and only if \( g \) is Frechet differentiable at \( x \), with \( \nabla g(x) = d(x) \).

P 3.5. Suppose that \( g_1, \ldots, g_N \in \mathbb{W}_0(\mathcal{E}) \) and \( \alpha_1, \ldots, \alpha_N \in \mathbb{R} \). Then we have for each \( x \in \mathcal{E} \),

\[
\partial \left\{ \sum_{j=1}^{N} \alpha_j g_j \right\}(x) \subseteq \sum_{j=1}^{N} \alpha_j \partial g_j(x)
\]

We will make the following assumptions regarding the function \( F : \mathcal{E} \times \Xi \to \mathbb{R} \).

A 3.1. \( F(x, \cdot) : \Xi \to \mathbb{R} \) is \( \mathcal{G} \)-measurable and \( \mathbb{Q} \)-integrable for each \( x \in \mathcal{E} \).

A 3.2. There exists a \( \mathcal{G} \)-measurable and \( \mathbb{Q} \)-integrable function \( K : \Xi \to \mathbb{R}_+ \), such that for \( \mathbb{Q} \)-almost all \( \zeta \), give any \( x, y \in \mathcal{E} \) we have

\[
|F(x, \zeta) - F(y, \zeta)| \leq K(\zeta) \|x - y\|_2
\]

In particular, there exists a \( \mathbb{Q} \)-null set \( N_\Xi \subset \Xi \) such that for \( \zeta \notin N_\Xi \), (15) holds for \( K(\zeta) < \infty \).

A 3.3. For any fixed \( x \in \mathcal{E} \), there exists a \( \mathbb{Q} \)-null set \( N_\Xi^x(x) \subset \Xi \) such that for \( \zeta \notin N_\Xi^x(x) \), \( F(\cdot, \zeta) \) is differentiable at \( x \), i.e., \( \nabla F(x, \zeta) \) exists. Consequently, from Property P 3.4, \( \partial F(x, \zeta) = \{\nabla F(x, \zeta)\} \).

Under these assumptions, we show next using results in ? and ?, that the sequence \( \{\hat{f}(\cdot, N)(\omega)\}_{N \in \mathbb{N}} \) (for \( \mathbb{P} \)-almost all \( \omega \in \Omega \)) converges to \( f \) in the Lipschitz norm.

Lemma 3.2. Suppose \( F : \mathcal{E} \times \Xi \to \mathbb{R} \) satisfies Assumptions A 3.1 through A 3.3. Then, given any compact set \( D \subset \mathcal{E} \), the following assertions hold.

1. \( f \) (as defined in (11)) is continuously differentiable on \( \mathcal{E} \), i.e., \( \nabla f(x) \) exists and is continuous for each \( x \in \mathcal{E} \). Further, \( f \in C_1(D) \).

2. For \( \mathbb{P} \)-almost all \( \omega \in \Omega \), \( \hat{f}(\cdot, N)(\omega) \in \mathbb{W}_0(D) \) for each \( N \in \mathbb{N} \)(where \( \hat{f} \) is as defined in (12)).

3. For \( \mathbb{P} \)-almost all \( \omega \in \Omega \), we have

\[
\lim_{N \to \infty} \sup_{x \in D} \left| \hat{f}(x, N)(\omega) - f(x) \right| = 0
\]
4. For $\mathbb{P}$-almost all $\omega \in \Omega$, we have

$$\lim_{N \to \infty} \sup_{x \in \mathcal{D}} \sup_{d \in \partial f(x,N)(\omega)} \|d - \nabla f(x)\|_2 = 0 \quad (17)$$

Proof. We show each of the above assertions in order.

1. Under Assumption A 3.1 through A 3.3, Lemma A2 (on page 21) in ? shows that $\nabla f(x)$ exists and is continuous for each $x \in \mathcal{E}$. Therefore, $f$ is also continuous on $\mathcal{E}$. Now, since $\mathcal{D} \subset \mathcal{E}$ is compact, it is easily seen that $f$ and $\nabla f$ are bounded on $\mathcal{D}$. Consequently, $\|f\|_{C_1(D)} < \infty$ and hence we get that $f \in C_1(D)$.

2. For any $N \in \mathbb{N}$, $\omega \in \Omega$ and $x, y \in \mathcal{D}$, we get from (12)

$$\left| \hat{f}(x, N)(\omega) - \hat{f}(y, N)(\omega) \right| = \left| \sum_{j=1}^{N} (F(x, \zeta^j(\omega)) - F(y, \zeta^j(\omega))) \right|$$

$$\leq \sum_{j=1}^{N} \left| F(x, \zeta^j(\omega)) - F(y, \zeta^j(\omega)) \right| \quad (18)$$

Consider the set

$$\mathcal{N}_1 := \bigcup_{j \in \mathbb{N}} (\zeta^j)^{-1}(\mathcal{N}_1^j) \quad (19)$$

where $\mathcal{N}_1^j$ is defined in Assumption A 3.2. It is clear that $\mathcal{N}_1$ is $\mathbb{P}$-null and we get $\{\zeta^j(\omega)\}_{j \in \mathbb{N}} \subset \Xi \setminus \mathcal{N}_1$, . Therefore, for $\omega \in \Omega \setminus \mathcal{N}_1$, using (15) in (18) we get that

$$\left| \hat{f}(x, N)(\omega) - \hat{f}(y, N)(\omega) \right| \leq \left( \sum_{j=1}^{N} \frac{K(\zeta^j(\omega))}{N} \right) \|x - y\|_2 \quad \text{for any } x, y \in \mathcal{D} \quad (20)$$

where

$$\left( \sum_{j=1}^{N} \frac{K(\zeta^j(\omega))}{N} \right) < \infty$$

Consequently, $\hat{f}(\cdot, N)(\omega)$ is continuous on $\mathcal{D}$ for $\omega \in \Omega \setminus \mathcal{N}_1$. Since $\mathcal{D}$ is compact, we get that

$$\sup_{x \in \mathcal{D}} |\hat{f}(x, N)(\omega)| < K_N(\omega) \quad \text{for some } K_N(\omega) < \infty$$

Therefore, we finally get for $\omega \in \Omega \setminus \mathcal{N}_1$

$$\left\| \hat{f}(\cdot, N)(\omega) \right\|_{W_0(\mathcal{E})} = \sup_{x \in \mathcal{D}} |\hat{f}(x, N)(\omega)| + \sup_{x, x+y \in \mathcal{D}, y \neq 0} \left| \frac{\hat{f}(x+y, N)(\omega) - \hat{f}(x, N)(\omega)}{\|y\|_2} \right|$$

$$\leq K_N(\omega) + \left( \sum_{j=1}^{N} \frac{K(\zeta^j(\omega))}{N} \right) < \infty$$

Hence $\{\hat{f}(\cdot, N)(\omega)\}_{N \in \mathbb{N}} \subset W_0(\mathcal{D})$ for $\mathbb{P}$-almost all $\omega \in \Omega$. 16
3. It is clear from (15) in Assumption A 3.2 that for $Q$-almost all $\zeta$, the function $F(\cdot, \zeta)$ is continuous on $D$. Now, consider some $x^* \in D$ such that $\mathbb{E}_Q[F(x^*, \zeta)] < \infty$. Such an $x^*$ exists from Assumption A 3.1. Since $D$ is compact, there exists $K_D < \infty$ such that $\sup_{x \in D} \|x - x^*\|_2 < K_D$. Therefore, we get using (15) that for $Q$-almost all $\zeta$,

$$|F(x, \zeta)| \leq |F(x^*, \zeta)| + K(\zeta) \|x - x^*\|_2$$

$$\leq |F(x^*, \zeta)| + K(\zeta)K_D$$

¿From Assumption A 3.1 we get that $|F(x^*, \zeta)|$ is $Q$-integrable and from Assumption A 3.2, we know that $K(\zeta)$ is $Q$-integrable. Therefore the family $\{|F(x, \zeta)| : x \in D\}$ is dominated by a $Q$-integrable function. Thus, using, Lemma A1 (on page 67) in ?, we get that (16) holds.

4. Finally, we show that (17) holds. Note that the statement of (17) is well-defined since we have already shown that $f \in \mathcal{C}_1(D) \subset \mathcal{C}_1(E)$ and that there exists a $P$-null set $N_{1Ω} \subset \Omega$ defined as in (19), such that for all $\omega \notin N_{1Ω}$, $\{\hat{f}(\cdot, N)(\omega)\}_{N \in \mathbb{N}} \subset \mathcal{W}_0(D)$.

In order to show (17), we define for any $\bar{x} \in D$ and $\bar{\delta} > 0$,

$$B(\bar{x}, \bar{\delta}) := \left\{ x \in D : \|x - \bar{x}\|_2 < \bar{\delta} \right\}$$

(21)

Now, let us fix a sequence $\{\bar{\delta}_k\}_{k \in \mathbb{N}}$ of positive real numbers with $\bar{\delta}_k \to 0$ as $k \to \infty$ and define for each $k \in \mathbb{N}$ a function $G_k : \Xi \to \mathbb{R}$ as

$$G_k^\xi(\zeta) := \sup_{x \in B(\bar{x}, \bar{\delta}_k)} \sup_{d \in \partial F(x, \zeta)} \|d - \nabla F(\bar{x}, \zeta)\|_2$$

(22)

For each $k \in \mathbb{N}$, we know that $B(\bar{x}, \bar{\delta}_k) \subset D \subset E$. Also, we know from Assumption A 3.2 that for $\zeta \notin N_{1\Xi}$ (where $Q(N_{1\Xi}) = 0$), (15) holds for $x, y \in E$ and consequently we have

$$\sup_{x, y \in D \atop x \neq y} \frac{|F(x, \zeta) - F(y, \zeta)|}{\|x - y\|_2} \leq K(\zeta) \leq \infty$$

Therefore, for each $x \in D$, we can define a a Clarke generalized gradient $\partial F(x, \zeta)$ as in (14). Further, from Assumption A 3.3, we get that there exists a $Q$-null set $N_{2\Xi}(\bar{x})$ such that for $\zeta \notin N_{2\Xi}(\bar{x})$, $F(\cdot, \zeta)$ is Frechet differentiable at $\bar{x}$, i.e, $\nabla F(\bar{x}, \zeta)$ exists. Thus, it is easily seen that $\{G_k^\xi\}_{k \in \mathbb{N}}$ is a sequence of random variables defined on $\Xi \setminus (N_{1\Xi} \cup N_{2\Xi}(\bar{x}))$.

¿From Assumption A 3.2 and Property P 3.1, we know that if $\zeta \notin N_{1\Xi}$, $\partial F(x, \zeta) = \{\nabla F(x, \zeta)\}$ for $x \in D \setminus U_D(\zeta)$ where $L(U_D(\zeta)) = 0$. Using this we show next that for each $\zeta \in \Xi \setminus$
\( (\mathcal{N}_E^1 \cup \mathcal{N}_E^2(\bar{x})) \) and \( k \in \mathbb{N} \),

\[
G_k^\tilde{\varepsilon}(\zeta) = \sup_{x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k) \setminus \mathcal{U}_F(\zeta)} \|\nabla F(x, \zeta) - \nabla F(\bar{x}, \zeta)\|_2
\]  

(23)

From (22) and Property P 3.4, it is easily seen that

\[
G_k^\tilde{\varepsilon}(\zeta) \geq \sup_{x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k) \setminus \mathcal{U}_F(\zeta)} \sup_{d \in \partial F(x, \zeta)} \|d - \nabla F(\bar{x}, \zeta)\|_2 = \sup_{x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k) \setminus \mathcal{U}_F(\zeta)} \|\nabla F(x, \zeta) - \nabla F(\bar{x}, \zeta)\|_2
\]

Therefore, we only have to show the converse inequality to prove (23) To do this, first we note that for any set \( C \subset \mathbb{R}^l \) and \( d^* \in \mathbb{R}^l \), we have

\[
\sup_{d \in C} \|d - d^*\|_2 = \sup_{x \in \text{conv}(C)} \|d - d^*\|_2
\]  

(24)

Let us fix \( \zeta \in \Xi \setminus (\mathcal{N}_E^1 \cup \mathcal{N}_E^2(\bar{x})) \) and \( k \in \mathbb{N} \). Then, for each \( x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k) \), using (24) and the definition of the Clarke generalized gradient in (14), we get

\[
\sup_{d \in \partial F(x, \zeta)} \|d - \nabla F(\bar{x}, \zeta)\|_2 = \sup \left\{ \|d - \nabla F(\bar{x}, \zeta)\|_2 : d = \lim_{x \to \bar{x}} \nabla F(x_j, \zeta), \{x_j\}_{j \in \mathbb{N}} \subset \mathcal{B}(\bar{x}, \tilde{\delta}_k) \setminus \mathcal{U}_D(\zeta) \right\}
\]  

(25)

Now, consider some \( d \in \mathbb{R}^l \) such that \( d = \lim_{x \to \bar{x}} \nabla F(x_j, \zeta) \) for some sequence \( \{x_j\}_{j \in \mathbb{N}} \subset \mathcal{B}(\bar{x}, \tilde{\delta}_k) \setminus \mathcal{U}_D(\zeta) \). We get for each \( j \in \mathbb{N} \),

\[
\|d - \nabla F(\bar{x}, \zeta)\|_2 \leq \|d - \nabla F(x_j, \zeta)\|_2 + \|\nabla F(x_j, \zeta) - \nabla F(\bar{x}, \zeta)\|_2
\]

\[
\leq \|d - \nabla F(x_j, \zeta)\|_2 + \sup_{x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k) \setminus \mathcal{U}_D(\zeta)} \|\nabla F(x, \zeta) - \nabla F(\bar{x}, \zeta)\|_2
\]

Taking limits as \( j \to \infty \) on both sides of the last inequality above and noting that \( \nabla F(x_j, \zeta) \to d \), we get that

\[
\|d - \nabla F(\bar{x}, \zeta)\|_2 \leq \sup_{x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k) \setminus \mathcal{U}_D(\zeta)} \|\nabla F(x, \zeta) - \nabla F(\bar{x}, \zeta)\|_2
\]

Consequently, using (25),

\[
\sup_{d \in \partial F(x, \zeta)} \|d - \nabla F(\bar{x}, \zeta)\|_2 \leq \sup_{x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k) \setminus \mathcal{U}_D(\zeta)} \|\nabla F(x, \zeta) - \nabla F(\bar{x}, \zeta)\|_2
\]

Since the above inequality is true for each \( x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k) \), we finally get

\[
G_k^\tilde{\varepsilon}(\zeta) := \sup_{x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k)} \sup_{d \in \partial F(x, \zeta)} \|d - \nabla F(\bar{x}, \zeta)\|_2 \leq \sup_{x \in \mathcal{B}(\bar{x}, \tilde{\delta}_k) \setminus \mathcal{U}_D(\zeta)} \|\nabla F(x, \zeta) - \nabla F(\bar{x}, \zeta)\|_2
\]

Thus, we have shown the converse inequality and hence (23) holds for each \( k \in \mathbb{N} \) and \( \zeta \in \Xi \setminus (\mathcal{N}_E^1 \cup \mathcal{N}_E^2(\bar{x})) \).

As a consequence, we get the following two properties of the sequence \( \{G_k^\tilde{\varepsilon}\}_{k \in \mathbb{N}} \).
• For each \( k \in \mathbb{N} \) and \( \zeta \in \Xi \setminus (\mathcal{N}^1_\Omega \cup \mathcal{N}^2_\Omega(\tilde{x})) \), we get from from Property P 3.2 that

\[
G^x_k(\zeta) \leq 2 \sup_{x \in B(\tilde{x}, \tilde{\delta}_k) \setminus \mathcal{U}_\zeta} \| \nabla F(x, \zeta) \|_2 \leq 2 \sup_{x \in \mathcal{D} \setminus \mathcal{U}_\zeta} \| \nabla F(x, \zeta) \|_2 \leq 2K(\zeta)
\]

Thus, for each \( k \in \mathbb{N} \), \( G^x_k \) is bounded above by the integrable function \( 2K(\zeta) \) for \( \mathcal{Q} \)-almost all \( \zeta \).

• For \( \zeta \in \Xi \setminus (\mathcal{N}^1_\Omega \cup \mathcal{N}^2_\Omega(\tilde{x})) \), we get Property P 3.3 and the fact that \( \tilde{\delta}_k \to 0 \) as \( k \to \infty \)

\[
\lim_{k \to \infty} G^x_k(\zeta) = \lim_{k \to \infty} \sup_{x \in B(\tilde{x}, \tilde{\delta}_k) \setminus \mathcal{U}_\zeta} \| \nabla F(x, \zeta) - \nabla F(\tilde{x}, \zeta) \|_2 = 0
\]

Thus, the sequence \( \{G^x_k\}_{k \in \mathbb{N}} \) converges point wise (in \( \zeta \)) to 0 for \( \mathcal{Q} \)-almost all \( \zeta \).

Therefore, using the Lebesgue Dominated Convergence Theorem, we get that

\[
\lim_{k \to \infty} \mathbb{E}_\mathcal{Q}[G^x_k(\zeta)] = 0
\]

That is, for each \( \epsilon > 0 \), there exists \( k(\tilde{x}, \epsilon) \in \mathbb{N} \), such that for all \( k \geq k(\tilde{x}, \epsilon) \), \( \mathbb{E}_\mathcal{Q}[G^x_k(\zeta)] < \epsilon \).

Now, we define the \( \mathcal{P} \)-null set \( \mathcal{N}^2_\Omega(\tilde{x}) \subset \Omega \) as

\[
\mathcal{N}^2_\Omega(\tilde{x}) := \bigcup_{j \in \mathbb{N}} (\zeta^j)^{-1}(\mathcal{N}^2_\mathbb{E}(\tilde{x})) \tag{26}
\]

If \( \omega \in \Omega \setminus \mathcal{N}^2_\Omega(\tilde{x}) \), then by definition \( \zeta^j(\omega) \notin \mathcal{N}^2_\mathbb{E}(\tilde{x}) \) for each \( j \in \mathbb{N} \). Recall that we also analogously defined the \( \mathcal{P} \)-null set \( \mathcal{N}^1_\Omega \subset \Omega \) in (19). Next, we consider the sample average functions \( \{\hat{f}(\cdot, N)(\omega)\}_{N \in \mathbb{N}} \) for \( \omega \in \Omega \setminus (\mathcal{N}^1_\Omega \cup \mathcal{N}^2_\Omega(\tilde{x})) \). First of all it is easily seen that for \( \omega \in \Omega \setminus (\mathcal{N}^1_\Omega \cup \mathcal{N}^2_\Omega(\tilde{x})) \), \( \hat{f}(\cdot, N)(\omega) \) is differentiable at \( \tilde{x} \) for each \( N \in \mathbb{N} \) and

\[
\nabla \hat{f}(\tilde{x}, N)(\omega) = \frac{1}{N} \sum_{j=1}^{N} \nabla F(\tilde{x}, \zeta^j(\omega))
\]

Further, for any \( x \in \mathcal{D} \), \( N \in \mathbb{N} \) and \( \omega \in \Omega \setminus (\mathcal{N}^1_\Omega \cup \mathcal{N}^2_\Omega(\tilde{x})) \), it is easily seen from Property P 3.5.

\[
\partial \hat{f}(x, N)(\omega) = \partial \left\{ \frac{1}{N} \sum_{j=1}^{N} F(\cdot, \zeta^j(\omega)) \right\}(x) \subseteq \frac{1}{N} \sum_{j=1}^{N} \partial F(x, \zeta^j(\omega))
\]
Using these two observations, we get for $\omega \in \Omega \setminus (\mathcal{N}_{\Omega}^1 \cup \mathcal{N}_{\Omega}^2(\bar{x}))$, $k \in \mathbb{N}$ and $x \in B(\bar{x}, \delta_k)$,

\[
\sup_{d \in \partial f(x,N)(\omega)} \left\| d - \nabla f(\bar{x}, N)(\omega) \right\|_2 = \sup \left\{ \left\| d - \nabla \hat{f}(\bar{x}, N)(\omega) \right\|_2 : d \in \partial \left\{ \frac{1}{N} \sum_{j=1}^{N} F(\cdot, \zeta^j(\omega)) \right\}(x) \right\}
\]
\[
\leq \sup \left\{ \left\| d - \nabla \hat{f}(\bar{x}, N)(\omega) \right\|_2 : d \in \frac{1}{N} \sum_{j=1}^{N} \partial F(x, \zeta^j(\omega)) \right\}
\]
\[
= \sup \left\{ \frac{1}{N} \left\| \sum_{j=1}^{N} (d^j - \nabla F(\bar{x}, \zeta^j(\omega))) \right\|_2 : d^j \in \partial F(x, \zeta^j(\omega)), \ j = 1, \ldots, N \right\}
\]
\[
\leq \sup \left\{ \frac{1}{N} \sum_{j=1}^{N} \left\| d^j - \nabla F(\bar{x}, \zeta^j(\omega)) \right\|_2 : d^j \in \partial F(x, \zeta^j(\omega)), \ j = 1, \ldots, N \right\}
\]
\[
= \frac{1}{N} \sum_{j=1}^{N} \sup_{d^j \in \partial F(x, \zeta^j(\omega))} \left\| d^j - \nabla F(\bar{x}, \zeta^j(\omega)) \right\|_2
\]

Therefore, we get that for $\omega \in \Omega \setminus (\mathcal{N}_{\Omega}^1 \cup \mathcal{N}_{\Omega}^2(\bar{x}))$ and $k \in \mathbb{N}$

\[
\sup_{x \in B(\bar{x}, \delta_k)} \sup_{d \in \partial f(x,N)(\omega)} \left\| d - \nabla f(\bar{x}, N)(\omega) \right\|_2 = \sup_{x \in B(\bar{x}, \delta_k)} \left\{ \frac{1}{N} \sum_{j=1}^{N} \sup_{d^j \in \partial F(x, \zeta^j(\omega))} \left\| d^j - \nabla F(\bar{x}, \zeta^j(\omega)) \right\|_2 \right\}
\]
\[
\leq \frac{1}{N} \sum_{j=1}^{N} \sup_{x \in B(\bar{x}, \delta_k)} \sup_{d^j \in \partial F(x, \zeta^j(\omega))} \left\| d^j - \nabla F(\bar{x}, \zeta^j(\omega)) \right\|_2
\]
\[
= \frac{1}{N} \sum_{j=1}^{N} G_k^x(\zeta^j(\omega))
\]

Now, consider the right side of the last inequality given above. Since $\{\zeta^j\}_{j \in \mathbb{N}}$ is an i.i.d. sequence, we get from the strong law large numbers that $\frac{1}{N} \sum_{j=1}^{N} G_k^x(\zeta^j)$ converges to $\mathbb{E}_x \left[ G_k^x(\zeta) \right] = \mathbb{E}_P \left[ G_k^x(\zeta^1(\omega)) \right]$ for $P$-almost all $\omega$. That is, there exists a $P$-null set $\mathcal{N}_{\Omega}^3(\bar{x}, k) \subset \Omega$ such that for $\omega \in \Omega \setminus (\mathcal{N}_{\Omega}^1 \cup \mathcal{N}_{\Omega}^2(\bar{x}) \cup \mathcal{N}_{\Omega}^3(\bar{x}, k))$, the following statement holds. For each $\epsilon > 0$, there exists $N(\bar{x}, k, \omega, \epsilon)$ such that

\[
\left| \sum_{j=1}^{N} G_k^x(\zeta^j(\omega)) \right| - \mathbb{E}_x \left[ G_k^x(\zeta) \right] < \epsilon \quad \text{for all} \quad N \geq N(\bar{x}, k, \omega, \epsilon)
\]

Therefore, we finally get that given any $\epsilon > 0$ and $\bar{x} \in \mathbb{D}$, for each $k > k(\bar{x}, \epsilon)$, there exists a $P$-null set $\mathcal{N}_{\Omega}^3(\bar{x}, k)$ such that if $\omega \in \Omega \setminus (\mathcal{N}_{\Omega}^1 \cup \mathcal{N}_{\Omega}^2(\bar{x}) \cup \mathcal{N}_{\Omega}^3(\bar{x}, k))$, then

\[
\sup_{x \in B(\bar{x}, \delta_k)} \sup_{d \in \partial f(x,N)(\omega)} \left\| d - \nabla \hat{f}(\bar{x}, N)(\omega) \right\|_2 < 2\epsilon \quad \text{for all} \quad N > N(\bar{x}, k, \omega, \epsilon) \quad (27)
\]

20
Also, we have already shown that \( f \in C_1(D) \). Since \( D \) is compact, this means that \( \nabla f \) is uniformly continuous on \( D \). That is, for each \( \varepsilon > 0 \), there exists \( \delta(\varepsilon) > 0 \) such that

\[
\sup\{\|\nabla f(x) - \nabla f(y)\|_2 : x, y \in D, \|y - x\|_2 \leq \delta(\varepsilon)\} < \varepsilon \quad (28)
\]

Now, consider the sequence \( \{(1/i)\}_{i \in \mathbb{N}} \) and some \( i \in \mathbb{N} \). For each \( \tilde{x} \in D \), we first pick \( k^i_x \in \mathbb{N} \) such that \( k^i_x > k(\tilde{x}, (1/i)) \) and \( \delta_{k^i_x} < \delta(1/i) \). Then clearly, the collection \( \{B(\tilde{x}, \delta_{k^i_x})\}_{\tilde{x} \in D} \) is an open cover for \( D \) (in the subspace topology on \( D \)). Since \( D \) is compact, there exists a finite sub-cover for \( D \); i.e.; there exist \( \tilde{x}_1, \ldots, \tilde{x}_m \in D \) such that

\[
D \subseteq \bigcup_{j=1}^{m} B(\tilde{x}_j, k^i_{\tilde{x}_j}) \quad (29)
\]

For each \( j = 1, \ldots, m \), the following statements clearly hold.

- From Assumption A 3.2, if \( \omega \in \Omega \setminus \mathcal{N}^2(\tilde{x}_j^i) \), then \( \nabla \hat{f}(\tilde{x}_j^i, N) \) exists for each \( N \in \mathbb{N} \).
- From the Strong Law of Large Numbers, there exists a \( \mathbb{P} \)-null set \( \mathcal{N}^4(\tilde{x}_j^i) \) such that if \( \omega \in \Omega \setminus \left( \mathcal{N}^2(\tilde{x}_j^i) \cup \mathcal{N}^3(\tilde{x}_j^i, k^i_{\tilde{x}_j}) \right) \), then for any \( \varepsilon > 0 \), there exists \( \tilde{N}(\tilde{x}_j^i, \omega, \varepsilon) \) such that

\[
\left\| \nabla \hat{f}(\tilde{x}_j^i, N) - \nabla f(\tilde{x}_j^i) \right\|_2 < \varepsilon \quad \text{for all} \ N > \tilde{N}(\tilde{x}_j^i, \omega, \varepsilon) \quad (30)
\]

Let us define

\[
\mathcal{N}_\Omega(i) := \bigcup_{j=1}^{m} \left\{ \mathcal{N}^2(\tilde{x}_j^i) \cup \mathcal{N}^3(\tilde{x}_j^i, k^i_{\tilde{x}_j}) \cup \mathcal{N}^4(\tilde{x}_j^i) \right\}
\]

\[
N(i, \tilde{\omega}) := \max_{j=1, \ldots, m} \max \left\{ N(\tilde{x}_j^i, k^i_{\tilde{x}_j}), N(\tilde{x}_j^i, (1/i)) \right\}
\]

It is easily seen that \( \mathbb{P}(\mathcal{N}_\Omega(i)) = 0 \).

Now, for any \( \omega \in \Omega \setminus \mathcal{N}_\Omega(i) \) and \( x \in D \), we know from (29) that there exists \( j \in \{1, \ldots, m\} \) such that \( x \in B(\tilde{x}_j^i, \delta_{\tilde{x}_j^i}) \). Therefore, for each \( N > N(i, \omega) \left( \geq N(\tilde{x}_j^i, k^i_{\tilde{x}_j}, \tilde{\omega}, (1/i)) \right) \), we get from (27) that

\[
\sup_{d \in \partial \hat{f}(x, N)} \left\| d - \nabla \hat{f}(\tilde{x}_j^i, N)(\omega) \right\|_2 < \frac{2}{i}
\]

Similarly, for each \( N > N(i, \omega) \left( \geq \tilde{N}(\tilde{x}_j^i, \omega, (1/i)) \right) \), we have from (30) that

\[
\left\| \nabla \hat{f}(\tilde{x}_j^i, N) - \nabla f(\tilde{x}_j^i) \right\|_2 < (1/i)
\]

Finally, since we chose \( k^i_{\tilde{x}_j} \) such that \( \delta_{k^i_{\tilde{x}_j}} < \delta(1/i) \), we get from (28) that

\[
\left\| \nabla f(x) - \nabla \hat{f}(\tilde{x}_j^i) \right\|_2 < \varepsilon
\]
Thus, combining these three observations, we get
\[
\sup_{d \in \partial f(x,N)(\omega)} \|d - \nabla f(x)\|_2 \leq \sup_{d \in \partial f(x,N)(\omega)} \left\{ \|d - \nabla \hat{f}(\bar{x}^i_j, N)(\omega)\|_2 + \|\nabla \hat{f}(\bar{x}^i_j, N)(\omega) - \nabla f(\bar{x}^i_j)\|_2 \right\}
\leq \frac{4}{i}
\]

Since the above inequality is true for each \(x \in D\), we get that for \(\omega \in \Omega \setminus \mathcal{N}_\Omega(i)\) and all \(N > N(i, \omega)\)
\[
\sup_{x \in D} \sup_{d \in \partial f(x,N)(\omega)} \|d - \nabla f(x)\|_2 \leq \frac{4}{i} \tag{31}
\]

Finally, if we set \(\mathcal{N}_\Omega := \bigcup \{\mathcal{N}_\Omega(i) : i \in \mathbb{N}\}\), it is clear that \(\mathbf{P}(\mathcal{N}_\Omega) = 0\). Then, for \(\omega \in \Omega \setminus \mathcal{N}_\Omega\) we get that for each \(i \in \mathbb{N}\), there exists \(N(i, \omega)\) such that for \(N > N(i, \omega)\), (31) holds. Therefore we get that for \(\mathbf{P}\)-almost all \(\omega\),
\[
\lim_{N \to \infty} \sup_{x \in D} \sup_{d \in \partial f(x,N)(\omega)} \|d - \nabla f(x)\|_2 = 0
\]

□

From Lemma 3.2, we get the following corollary.

**Corollary 3.3.** Let Assumptions A 3.1 through A 3.3 hold. Then, given any compact set \(D \subset \mathcal{E}\) we have for \(\mathbf{P}\) almost all \(\omega\),
\[
\lim_{N \to \infty} \left\| \hat{f}(\cdot, N)(\omega) - f \right\|_{W_0(D)} = 0
\]

Consequently, there exists \(K_f(\omega) < \infty\), such that \(\left\| \hat{f}(\cdot, N)(\omega) \right\|_{W_0(D)} < K_f(\omega)\) and \(\|f\|_{W_0(D)} < K_f(\omega)\) for each \(N \in \mathbb{N}\).

**Proof.** Since all the assumptions of Lemma 3.2 hold, we see \(\{\hat{f}(\cdot, N)(\omega)\}_{N \in \mathbb{N}} \subset W_0(D)\) for \(\mathbf{P}\)-almost all \(\omega\). Also, \(f \in C_1(D) \subset W_0(D)\). Therefore, the statement of Corollary 3.3 is well defined.

Now, from Lemma 3.2, there exists a set \(\mathcal{N}_\Omega \subset \Omega\), such that for all \(\omega \in \Omega \setminus \mathcal{N}_\Omega\), \(\{\hat{f}(\cdot, N)(\omega)\}_{n \in \mathbb{N}} \subset W_0(D)\) and (16), (17) hold. Consider any such \(\omega \in \Omega \setminus \mathcal{N}_\Omega\). From Rademacher’s theorem, we know that for each \(N \in \mathbb{N}\), there exists a set \(\mathcal{U}_N(\omega) \subset D\) with \(\mathbf{L}(\mathcal{U}_N(\omega)) = 0\), such that \(\nabla \hat{f}(\cdot, N)(\omega)\) exists for all \(x \in D \setminus \mathcal{U}_N(\omega)\). Using Property P 3.4, we get that
\[
\left\| \hat{f}(\cdot, N)(\omega) - f \right\|_{W_0(D)} = \sup_{x \in D} |\hat{f}(x, N)(\omega) - f(x)| + \sup_{x \in D \setminus \mathcal{U}_N(\omega)} \left\| \nabla \hat{f}(x, N)(\omega) - \nabla f(x) \right\|_2 \tag{32}
\]
It follows from (16) that
\[
\lim_{N \to \infty} \sup_{x \in D} |\hat{f}(x, N)(\omega) - f(x)| = 0
\]
Also, for each \(N \in \mathbb{N}\),
\[
\sup_{x \in D} \sup_{d \in \partial \hat{f}(x, N)(\omega)} \|d - \nabla f(x)\|_2 \geq \sup_{x \in D \setminus \mathcal{U}_N(\omega)} \sup_{d \in \partial \hat{f}(x, N)(\omega)} \|d - \nabla f(x)\|_2
\]
But we know that for \(x \in D \setminus \mathcal{U}_N(\omega)\), \(\hat{f}(\cdot, N)(\omega)\) is differentiable and hence, \(\partial \hat{f}(\cdot, N) = \{\nabla \hat{f}(\cdot, N)\}\). Thus,
\[
\sup_{x \in D \setminus \mathcal{U}_N(\omega)} \sup_{d \in \partial \hat{f}(x, N)(\omega)} \|d - \nabla f(x)\|_2 = \sup_{x \in D \setminus \mathcal{U}_N(\omega)} \|\nabla \hat{f}(x, N)(\omega) - \nabla f(x)\|_2
\]
Therefore, we get from (17) that
\[
0 = \lim_{N \to \infty} \sup_{x \in D} \sup_{d \in \partial \hat{f}(x, N)(\omega)} \|d - \nabla f(x)\|_2 \geq \lim_{N \to \infty} \sup_{x \in D \setminus \mathcal{U}_N(\omega)} \|\nabla \hat{f}(x, N)(\omega) - \nabla f(x)\|_2
\]
Thus, we have for all \(\omega \in \Omega \setminus \mathcal{N}_\Omega\),
\[
\lim_{N \to \infty} \left\|\hat{f}(\cdot, N)(\omega) - f\right\|_{W_0(D)} = \lim_{N \to \infty} \sup_{x \in D} |\hat{f}(x, N)(\omega) - f(x)| + \lim_{N \to \infty} \sup_{x \in D \setminus \mathcal{U}_N(\omega)} \|\nabla \hat{f}(x, N)(\omega) - \nabla f(x)\|_2 = 0
\]
Therefore, \(\left\|\hat{f}(\cdot, N)(\omega) - f\right\|_{W_0(D)} \to 0\) as \(N \to \infty\) for \(P\)-almost all \(\omega\). Hence \(\left\|\hat{f}(\cdot, N)(\omega)\right\|_{W_0(D)} \to \|f\|_{W_0(D)}\) as \(N \to \infty\) and consequently, the sequence \(\{\left\|\hat{f}(\cdot, N)(\omega)\right\|_{W_0(D)}\}_{N \in \mathbb{N}}\) is bounded. Therefore, there exists \(K_f(\omega) \in (\|f\|_{W_0(D)}, \infty)\) such that \(\left\|\hat{f}(\cdot, N)(\omega)\right\|_{W_0(D)} < K_f\) for each \(N \in \mathbb{N}\). \(\square\)

In the rest of this article we will assume that we have an approximating function \(\hat{f}(\cdot, N)\) that converges to our objective function \(f\) as the approximation parameter \(N \to \infty\) in the appropriate sense. If \(f\) is an expected value function as in (11)we understand that the convergence of \(\hat{f}(\cdot, N)(\omega)\) to \(f\) is in the almost-sure sense the sample size \(N\) (which is the approximation parameter in this case) goes to \(\infty\). In particular, we assume that \(\tilde{\omega} \in \Omega\) has been appropriately chosen to ensure the convergence of \(\hat{f}(\cdot, N)(\tilde{\omega})\) to \(f\) and therefore, we drop the dependence of \(\hat{f}\) on \(\omega\) and denote the sample average function as \(\hat{f}(\cdot, N)\).

### 4 Local Approximations of the Objective Function

In the previous section, we looked at the construction of a function \(\hat{f}(\cdot, N)\) that approximates an objective function \(f\) as defined in (11) and we showed the convergence of \(\hat{f}(\cdot, N)\) to \(f\) as the
approximation parameter (the sample size in this case) \( N \to \infty \). Now, in this section, we assume the existence of an approximation \( \hat{f}(\cdot, N) \) for \( f \) and show how \( \hat{f} \) can be used to solve the optimization problem \((P)\).

When the objective function \( f \) and its higher order derivatives are easily evaluated, a typical algorithm used to solve the problem \((P)\) would work as follows. We would start at some \( x_0 \in \mathcal{X} \) and generate a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that the limit points of this sequence satisfy an appropriate optimality condition. That is, at each \( x_n \in \mathcal{X} \) we would use the function value \( f(x_n) \), the gradient \( \nabla f(x_n) \) and possibly the Hessian \( \nabla^2 f(x_n) \) (if \( f \) is sufficiently smooth) to find the next point \( x_{n+1} \in \mathcal{X} \) in the sequence. However, we are interested in cases where \( f \) and its higher order derivatives cannot be evaluated and \( f \) can only be approximated by the function \( \hat{f} \). Accordingly, in this section, we consider techniques to generate sequences \( \{f_n : E \mapsto \mathbb{R}\}_{n \in \mathbb{N}}, \{\hat{\nabla}_n f : E \mapsto \mathbb{R}^l\}_{n \in \mathbb{N}} \) and \( \{\hat{\nabla}_n^2 f : E \mapsto S\}_{n \in \mathbb{N}} \) converging respectively to \( f \), \( \nabla f \) and \( \nabla^2 f \) as \( n \to \infty \). The idea as before, is to start at some \( x_0 \in \mathcal{X} \) and for each \( n \in \mathbb{N} \), use \( f_n(x_n), \hat{\nabla}_n f(x_n) \) and \( \hat{\nabla}_n^2 f(x_n) \) in place of \( f(x_n), \nabla f(x_n) \) and \( \nabla^2 f(x_n) \) respectively.

Let us begin by constructing the sequence \( \{f_n\}_{n \in \mathbb{N}} \). Let us define a sequence \( \{N^0_n\}_{n \in \mathbb{N}} \subset \mathbb{N} \) and set for each \( x \in E \) and \( n \in \mathbb{N} \).

\[
\begin{align*}
&f_n(x) := \hat{f}(x, N^0_n) \quad (33)
\end{align*}
\]

**Lemma 4.1.** Suppose the following assumptions hold.

**A 4.1.** For any compact set \( D \subset E \), \( f \) is continuous on \( D \) and the sequence \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \) converges uniformly to \( f \) on \( D \).

\[
\lim_{N \to \infty} \sup_{x \in D} |\hat{f}(x, N) - f(x)| = 0 \quad (34)
\]

**A 4.2.** The sequence \( \{N^0_n\}_{n \in \mathbb{N}} \) of sample sizes satisfies \( N^0_n \to \infty \) as \( n \to \infty \).

Then, for any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset D \subset E \), we get

\[
\begin{align*}
\lim_{n \to \infty} |f_n(x_n) - f(x_n)| &= 0
\end{align*}
\]

In particular, if \( x_n \to \bar{x} \), then

\[
\begin{align*}
\lim_{n \to \infty} f_n(x_n) &= f(\bar{x})
\end{align*}
\]

**Proof.** We have for each \( n \in \mathbb{N} \).

\[
|f_n(x_n) - f(x_n)| = |\hat{f}(x, N^0_n) - f(x_n)| \leq \sup_{x \in D} |\hat{f}(x, N^0_n) - f(x)|
\]

24
Therefore, taking limits as \( n \to \infty \) and noting that \( N^0_n \to \infty \), we get
\[
\lim_{n \to \infty} |f_n(x_n) - f(x_n)| \leq \limsup_{n \to \infty} \left| \hat{f}(x, N^0_n) - f(x) \right| = 0
\]

Next, from the continuity of \( f \) on \( D \), we get
\[
\lim_{n \to \infty} |f_n(x_n) - f(\tilde{x})| \leq \lim_{n \to \infty} |f_n(x_n) - f(x_n)| + \lim_{n \to \infty} |f(x_n) - f(\tilde{x})| = 0
\]

Thus, we see that if we have uniform convergence of the sequence \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \) of approximating functions to the function \( f \) on some compact set \( D \supset \{x_n\}_{n \in \mathbb{N}} \), then the accuracy of \( f_n(x_n) \) as an approximation to \( f(x_n) \), can be improved by increasing the sample size \( N^0_n \) to \( \infty \) as \( n \to \infty \).

Remark : Note that for the case when \( f \) is an expected value function as in (11) and \( \hat{f}(\cdot, N)(\omega) \) is the sample average function, we can modify Assumption A 4.2 and require that \( N^0_n \to \infty \) for \( \mathbb{P} \)-almost all \( \omega \). With such a condition, it is easy to see from Lemma 4.1 how we can get \( \mathbb{P} \)-almost sure convergence of \( f_n(x_n) \) to \( f(\tilde{x}) \).

Next let us consider how the sequences \( \{\hat{\nabla}_n f\}_{n \in \mathbb{N}} \) and \( \{\hat{\nabla}^2_n f\}_{n \in \mathbb{N}} \) can be generated. Our approach is as follows. As before, let us assume that we have a sequence \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \) converging to \( f \) in an appropriate sense as \( N \to \infty \). Given any \( n \in \mathbb{N} \) and point \( x \in \mathcal{E} \), we evaluate \( \hat{f}(x + y_i^n, N^n) \) for \( i = 1, \ldots, M_n \) at selected points \( x + y_1^n, \ldots, x + y_{M_n}^n \) in a neighborhood of \( x \) with appropriate approximation parameter values \( \{N^n_1, \ldots, N^n_{M_n}\} \) at each such point. Then we use weighted regression to fit a linear or quadratic function \( \hat{f}_n \) to these observations. Thus, \( \hat{f}_n \) is a linear or quadratic model that locally approximates the function \( f \) in a neighborhood of \( x \). Then, the gradient \( \nabla \hat{f}_n(x) \) and the Hessian \( \nabla^2 \hat{f}_n(x) \) of the local approximation \( \hat{f}_n \) give approximations of the gradient \( \nabla f(x) \) and the Hessian \( \nabla^2 f(x) \) respectively of \( f \) at the point \( x \). We use \( \hat{\nabla}_n f(x) \) and \( \hat{\nabla}^2_n f(x) \) to denote generic approximations of the gradient \( \nabla f(x) \) and the Hessian \( \nabla^2 f(x) \) respectively of \( f \) at the point \( x \), and we use \( \nabla \hat{f}_n(x) \) and \( \nabla^2 \hat{f}_n(x) \) to denote the gradient and the Hessian of the local approximation \( \hat{f}_n \) at \( x \). In specific cases we will choose \( \hat{\nabla}_n f(x) = \nabla \hat{f}_n(x) \) and \( \hat{\nabla}^2_n f(x) = \nabla^2 \hat{f}_n(x) \).

Our convergence results regarding the sequences \( \{\nabla \hat{f}_n\}_{n \in \mathbb{N}} \) and \( \{\nabla^2 \hat{f}_n\}_{n \in \mathbb{N}} \) are as follows. Consider a compact set \( D \subset \mathcal{E} \) and a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset D \). Assume that \( f \in \mathcal{C}_1(\mathcal{E}) \), \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \subset \mathcal{W}_0(D) \) and \( \left\| \hat{f}(\cdot, N) - f \right\|_{\mathcal{W}_0(D)} \to 0 \) as \( N \to \infty \) for every compact set \( D \subset \mathcal{E} \). Suppose for each
n ∈ N we fit a linear or quadratic model \( \hat{f}_n \) using regression and determine \( \nabla \hat{f}_n(x_n) \) and \( \nabla^2 \hat{f}_n(x_n) \).

Then, under certain conditions on the design points used in the regression for each \( n \in N \) and the corresponding values of the approximation parameter \( N_i^n \) for each of design point, we show that

\[
\left\| \nabla \hat{f}_n(x_n) - \nabla f(x_n) \right\|_2 \to 0 \quad \text{as} \quad n \to \infty.
\]

Note that we can use a quadratic model \( \hat{f}_n \) to approximate \( f \) even if \( \nabla^2 f(x) \) does not exist at all \( x \). Further, if \( f \in C^2(E) \), \( \{ \hat{f}(\cdot, N) \}_{N \in N} \subset W^1(D), \)

\[
\left\| \hat{f}(\cdot, N) - f \right\|_{W^1(D)} \to 0 \quad \text{as} \quad N \to \infty \quad \text{for} \quad \text{every compact set} \quad D \subset E \quad \text{and} \quad \hat{f}_n \quad \text{is a quadratic model for each} \quad n \in N \quad \text{then under some additional conditions on the design points of the regression for each} \quad n \in N, \quad \text{we show the convergence of both} \quad \left\| \nabla \hat{f}_n(x_n) - \nabla f(x_n) \right\|_2 \quad \text{and} \quad \left\| \nabla^2 \hat{f}_n(x_n) - \nabla^2 f(x_n) \right\|_2 \to 0 \quad \text{as} \quad n \to \infty. \]

It is important to note that in either case we do not rely on exact evaluations of \( f(x) \) or \( \nabla f(x) \) or \( \nabla^2 f(x) \) at any \( x \in E \) in order to obtain convergence. We also do not rely on exact evaluations or even the existence of \( \nabla f_n(x) \) or \( \nabla^2 f_n(x) \) at any \( x \in E \). Instead, we use only the values of the approximating function \( \hat{f} \) at selected points using selected approximation parameters.

In order to provide some background, suppose that \( f \) can be evaluated exactly at any \( x \in E \), and consider the well known method of finite differences to construct a gradient approximation, and thus a first order local approximation \( \hat{f}_n \) of \( f \) in a neighborhood of \( x \in E \). The function \( f \) is evaluated at \( l + 1 \) points \( \{ x, x + c_n e_1, \ldots, x + c_n e_l \} \), where \( c_n > 0 \) and \( \{ e_1, \ldots, e_l \} \) is the standard basis for \( \mathbb{R}^l \). Then the local approximation \( \hat{f}_n \) is given by

\[
\hat{f}_n(y) := f(x) + \nabla \hat{f}_n(x)^T (y - x) \quad \text{for all} \quad y \in \mathbb{R}^l
\]

where \( \nabla \hat{f}_n(x) \) is given by

\[
\nabla \hat{f}_n(x) := \begin{pmatrix}
f(x + c_n e_1) - f(x) \\
c_n \\
f(x + c_n e_2) - f(x) \\
c_n \\
\vdots \\
f(x + c_n e_l) - f(x) \\
c_n
\end{pmatrix}
\]

The following convergence result for the sequence \( \{ \nabla \hat{f}_n(x) \}_{n \in N} \) is well known.

**Theorem 4.2.** Consider an open set \( E \) and any \( x \in E \). Suppose that the following assumptions hold.

1. The function \( f : E \to \mathbb{R} \) is continuously differentiable on \( E \).

2. The gradient approximation \( \nabla \hat{f}_n(x) \) is given by (35) for each \( n \in N \).
3. The step sizes $c_n \to 0$ as $n \to \infty$.

Then, for each $x \in E$, $\nabla \hat{f}_n(x) \to \nabla f(x)$ as $n \to \infty$.

Proof. Recall that

$$\nabla f(x) := \begin{pmatrix}
\frac{\partial f}{\partial x_1}(x) \\
\vdots \\
\frac{\partial f}{\partial x_l}(x)
\end{pmatrix}$$

where

$$\frac{\partial f}{\partial x_i}(x) := \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h} \quad \text{for each } i \in \{1, \ldots, l\}$$  \hspace{1cm} (36)

The limit in (36) exists for each $i \in \{1, \ldots, l\}$ since $f$ is continuously differentiable on $E$. Since $c_n \to 0$ as $n \to \infty$,

$$\lim_{n \to \infty} \frac{f(x + c_ne_i) - f(x)}{c_n} = \frac{\partial f}{\partial x_i}(x) \quad \text{for each } i \in \{1, \ldots, l\}$$

Therefore, $\nabla \hat{f}_n(x) \to \nabla f(x)$ as $n \to \infty$. \hfill \square

We generalize the method described above in the following manner. Given any $n \in \mathbb{N}$ and $x \in E$, we consider a set of points in the neighborhood of $x$ points be denoted by $\{x + y_1^n, \ldots, x + y_{M_n}^n\}$, the corresponding set of approximation parameter values be given by $N_n := \{N_1^n, \ldots, N_{M_n}^n\}$ and the corresponding set of non-negative weights be given by $\{w_1^n, \ldots, w_{M_n}^n\}$. We evaluate $\hat{f}(x + y_i^n, N_i^n)$ for $i = 1, \ldots, M_n$ and $\hat{f}(x, N^n_i)$ for each $i = 0, \ldots, M_n$. Then, our first order local approximation $\hat{f}_n$ is given by

$$\hat{f}_n(y) := \hat{f}(x, N^n_0) + \nabla \hat{f}_n(x)^T(y - x)$$  \hspace{1cm} (37)

where $\nabla \hat{f}_n(x)$ is given by

$$\nabla \hat{f}_n(x) \in \arg \min_{\beta \in \mathbb{R}^l} \sum_{i=1}^{M_n} \left[ w_i^n \left\{ \hat{f}(x + y_i^n, N_i^n) - \hat{f}(x, N_i^n) - \beta^T y_i^n \right\}^2 \right]$$  \hspace{1cm} (38)

Under some conditions, it is possible to give a closed-form expression for $\nabla \hat{f}_n(x)$, as in (35). Next, we introduce notation to do so. Let the weight matrix be give by $W_n := \text{diag}(w_1^n, \ldots, w_{M_n}^n)$ and the perturbation matrix $Y_n \in \mathbb{R}^{M_n \times l}$ be given by

$$Y_n := \begin{pmatrix}
(y_1^n)^T \\
(y_2^n)^T \\
\vdots \\
(y_{M_n}^n)^T
\end{pmatrix}$$  \hspace{1cm} (39)

27
For any \( x \in \mathcal{E} \), perturbation matrix \( Y_n \) and set of approximation parameters \( N_n \) we define the column vector \( \hat{f}(x, Y_n, N_n) \in \mathbb{R}^{M_n} \).

\[
\hat{f}(x, Y_n, N_n) := \begin{pmatrix}
\hat{f}(x + y^2_n, N^2_n) - \hat{f}(x, N^2_n) \\
\hat{f}(x + y^1_n, N^1_n) - \hat{f}(x, N^1_n) \\
\vdots \\
\hat{f}(x + y^{M_n}_n, N^{M_n}_n) - \hat{f}(x, N^{M_n}_n)
\end{pmatrix}
\]

(40)

Similarly, we define \( f(x, Y_n) \in \mathbb{R}^{M_n} \) for any \( x \in \mathbb{R}^l \) as

\[
f(x, Y_n) := \begin{pmatrix}
f(x + y^1_n) - f(x) \\
f(x + y^2_n) - f(x) \\
\vdots \\
f(x + y^{M_n}_n) - f(x)
\end{pmatrix}
\]

(41)

Then, it is easy to show that

\[
\arg \min_{\beta \in \mathbb{R}^l} \left\{ \sum_{i=1}^{M_n} w^i_n \left[ f_n(x + y^i_n) - f_n(x) - \beta^T y^i_n \right]^2 \right\} = \left\{ \beta \in \mathbb{R}^l : (Y_n^T W_n Y_n) \beta = Y_n^T W_n \hat{f}(x, Y_n, N_n) \right\}
\]

(42)

If \( Y_n^T W_n Y_n \) is positive definite, then the unique optimal solution in (42) is given by

\[
\nabla \hat{f}_n(x_n) = (Y_n^T W_n Y_n)^{-1} Y_n^T W_n \hat{f}(x, Y_n, N_n)
\]

(43)

Note that we require \( M_n \geq l \) for \( Y_n^T W_n Y_n \) to be positive definite. Also note that the finite difference approximation is a special case of this approach with \( y^i_n = c_n e_i \) and \( w^i_n = 1 \) for each \( i \in \{1, \ldots, l\} \).

Next, suppose that we want to construct a quadratic local approximation of \( f \) in a neighborhood of \( x \in \mathcal{E} \) for some \( n \in \mathbb{N} \). Again, we evaluate \( \hat{f}(x + y^i_n, N^i_n) \) for \( i = 1, \ldots, M_n \) and \( \hat{f}(x, N^i_n) \) for \( i = 0, \ldots, M_n \). Then, our second order local approximation \( \hat{f}_n \) is given by

\[
\hat{f}_n(y) := \hat{f}(x, N^0_n) + \nabla \hat{f}_n(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 \hat{f}_n(x) (y - x)
\]

(44)

where \( \nabla \hat{f}_n(x) \) and \( \nabla^2 \hat{f}_n(x) \) are given by

\[
(\nabla \hat{f}_n(x_n), \nabla^2 \hat{f}_n(x_n)) \in \arg \min_{(\beta, \Lambda) \in \mathbb{R}^l \times \mathbb{S}^{l \times l}} \sum_{i=1}^{M_n} w^i_n \left[ f_n(x_n + y^i_n) - f_n(x_n) - \beta^T y^i_n - \frac{1}{2} (y^i_n)^T \Lambda y^i_n \right]^2
\]

(45)

Note that if \( f \) is assumed to be twice continuously differentiable, then the Hessian matrix \( \nabla^2 f(x) \) is symmetric for all \( x \in \mathcal{X} \). Therefore, we choose \( \nabla^2 \hat{f}_n(x) \) also to be symmetric for each \( n \in \mathbb{N} \).
and $x \in \mathcal{E}$. Thus, to determine the second order approximation $\hat{f}_n$, we have to compute the $l$ components of $\nabla \hat{f}_n(x)$ and the $l(l+1)/2$ components of $\nabla^2 \hat{f}_n(x)$.

As in the case of the linear model, it is possible to find a closed form expression for $\nabla \hat{f}_n(x)$ and $\nabla^2 \hat{f}_n(x)$ in terms of appropriate regression matrices. To this end, we define the following notation. Consider any vector $y \in \mathbb{R}^l$ and matrix $H \in \mathbb{S}^{l \times l}$,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1l} \\ h_{21} & h_{22} & \cdots & h_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ h_{l1} & h_{l2} & \cdots & h_{ll} \end{pmatrix}$$

Then, the quadratic form $y^T Hy$ expands as follows:

$$y^T Hy = \sum_{j=1}^l \sum_{k=1}^l h_{jk} y_j y_k$$

Since $H$ is assumed to be symmetric we know that $h_{jk} = h_{kj}$ for all $j, k \in \{1, \ldots, l\}$. Thus,

$$y^T Hy = \sum_{j=1}^l h_{jj} y_j^2 + 2 \sum_{j=1}^{l-1} \sum_{k=j+1}^l h_{jk} y_j y_k$$

Therefore,

$$\frac{1}{2} y^T Hy = \frac{1}{2} \sum_{j=1}^l h_{jj} y_j^2 + \sum_{j=1}^{l-1} \sum_{k=j+1}^l h_{jk} y_j y_k$$

We wish to write the right side of the equation above as the scalar product of two appropriately defined vectors. Hence, corresponding to vector $y \in \mathbb{R}^l$, we define the vector $y^Q \in \mathbb{R}^{(l+1)/2}$ to be

$$y^Q := \left( \frac{1}{\sqrt{2}} y_1^2, \ldots, \frac{1}{\sqrt{2}} y_l^2, y_1 y_2, y_1 y_3, y_2 y_3, \ldots, y_1 y_l, y_2 y_l, \ldots, y_{l-1} y_l \right)^T$$

$$= \left( \{ \frac{1}{\sqrt{2}} y_j^2 \}_{j=1, \ldots, l}, \{ y_j y_k \}_{j,k=1, \ldots, l \atop j<k} \right)^T$$

Note the following useful relationship between the Euclidean norms of $y$ and $y^Q$:

$$\|y^Q\|_2^2 = \frac{1}{2} \left[ \sum_{j=1}^l y_j^4 + \sum_{j=1}^{l-1} \sum_{k=j+1}^l 2y_j^2 y_k^2 \right] = \frac{1}{2} \left( \sum_{j=1}^l y_j^2 \right)^2 = \frac{1}{2} \|y\|_4^2$$

(47)
Similarly, we write the components of \( H \) as a vector \( H_v \in \mathbb{R}^{l(l+1)/2} \) as follows

\[
H_v := \left( \frac{1}{\sqrt{2}}h_{11}, \ldots, \frac{1}{\sqrt{2}}h_{1l}, h_{12}, h_{13}, h_{23}, \ldots, h_{1l}, h_{2l}, \ldots, h_{l-1,l} \right)^T
\]

\[
= \left( \left\{ \frac{1}{\sqrt{2}}h_{jj} \right\}_{j=1,\ldots,l}, \left\{ h_{jk} \right\}_{j,k=1,\ldots,l, j<k} \right)^T
\]

Then,

\[
\frac{1}{2} y^T H y = H_v^T g^Q
\]

The following relationship between \( \|H_v\|_2 \) and \( \|H\|_F \) is also evident.

\[
\|H\|_F = \sqrt{\sum_{j=1}^{l} h_{jj}^2 + 2 \sum_{j=1}^{l-1} \sum_{k=j+1}^{l} h_{jk}^2}
\]

\[
= \sqrt{2} \sqrt{\sum_{j=1}^{l} \left( \frac{h_{jj}}{\sqrt{2}} \right)^2 + \sum_{j=1}^{l-1} \sum_{k=j+1}^{l} h_{jk}^2}
\]

\[
= \sqrt{2} \|H_v\|_2
\]

Now, we have the notation to write down the regression matrix for the quadratic model. Using the notation in (46), for each \( i \in \{1, \ldots, M_n\} \), let

\[
y^i_n := \left( \begin{array}{c}
\frac{1}{\sqrt{2}}(y^i_{n1})^2 \\
\vdots \\
\frac{1}{\sqrt{2}}(y^i_{n1})^2 \\
(y^i_n)_1(y^i_n)_2 \\
(y^i_n)_1(y^i_n)_3 \\
(y^i_n)_2(y^i_n)_3 \\
\vdots \\
(y^i_n)_1(y^i_n)_t \\
(y^i_n)_2(y^i_n)_t \\
\vdots \\
(y^i_n)_{t-1}(y^i_n)_t \\
\end{array} \right)
\]

30
where \((y_n^i)_j\) denotes the \(j^{th}\) component of \(y_n^i \in \mathbb{R}^l\). Also, define the matrix \(Y_n^Q \in \mathbb{R}^{M_n \times (l+1)/2}\) as

\[
Y_n^Q := \begin{pmatrix}
(y_n^1)^T \\
(y_n^2)^T \\
\vdots \\
(y_n^{M_n})^T
\end{pmatrix}
\]

(51)

Next, define the regression matrix \(Z_n \in \mathbb{R}^{M_n \times (l+1)(l+1)/2}\) for the second order approximation \(\hat{f}_n\) by

\[
Z_n = \begin{pmatrix}
(z_n^1)^T \\
\vdots \\
(z_n^{M_n})^T
\end{pmatrix} := \begin{pmatrix}
(y_n^T) & (y_n^1)^T \\
\vdots & \vdots \\
(y_n^{M_n})^T & (y_n^{M_n})^T
\end{pmatrix} = \begin{pmatrix} Y_n & Y_n^Q \end{pmatrix}
\]

(52)

Finally, we define \([\nabla^2 \hat{f}_n(x)]_v \in \mathbb{R}^{(l+1)/2}\) corresponding to \(\nabla^2 \hat{f}_n(x)\) by

\[
[\nabla^2 \hat{f}_n(x)]_v := \left\{ \frac{1}{\sqrt{2}} [\nabla^2 \hat{f}_n(x)]_{jj} \right\}_{j=1,\ldots,l} \cup \left\{ [\nabla^2 \hat{f}_n(x)]_{jk} \right\}_{j,k=1,\ldots,l, j<k}
\]

(53)

where \([\nabla^2 \hat{f}_n(x)]_{jk}\) denotes the element in the \(j^{th}\) row and \(k^{th}\) column of \(\nabla^2 \hat{f}_n(x)\).

Now, similar to the linear model, with the regression matrix \(Z_n\) defined as in (48), it can be shown that

\[
\arg \min_{(\beta, \Lambda) \in \mathbb{R}^{l} \times S^{l \times l}} \sum_{i=1}^{M_n} w_i \left[ f_n(x_n + y_n^i) - f_n(x_n) - \beta^T y_n^i - \frac{1}{2}(y_n^i)^T \Lambda y_n^i \right]^2 = \\
\left\{ \beta \in \mathbb{R}^l, \Lambda \in S^{l \times l} : Z_n^T W_n Z_n \begin{pmatrix} \beta \\ \Lambda \end{pmatrix} = Z_n^T W_n \hat{f}(x_n, Y_n, N_n) \right\}
\]

(54)

where the vector \(\Lambda_v \in \mathbb{R}^{l(l+1)/2}\) is the vector obtained by writing the matrix \(\Lambda \in S^{l \times l}\) in vector form as in (48). If \(Z_n^T W_n Z_n\) is positive definite, there is a unique element in the set (54) given by

\[
\begin{pmatrix}
\nabla \hat{f}_n(x_n) \\
[\nabla^2 \hat{f}_n(x_n)]_v
\end{pmatrix} = \left( Z_n^T W_n Z_n \right)^{-1} Z_n^T W_n \hat{f}(x_n, Y_n, N_n)
\]

(55)

It is easy to see that in this case, in order for \(Z_n^T W_n Z_n\) to be positive definite, it is necessary that \(M_n \geq l + l(l + 1)/2\).

It is also interesting to consider the case when \(Z_n^T W_n Z_n\) is a singular matrix and hence \(\nabla \hat{f}_n(x)\) and \(\nabla^2 \hat{f}_n(x)\) cannot be determined by (55). However, the set of optimal solutions (54) still is
non-empty in this case and any one of its elements can be chosen as $\nabla \hat{f}_n(x)$ and $\nabla^2 \hat{f}_n(x)$. Thus, irrespective of whether $Z_n^T W_n Z_n$ is positive definite or not, we can find $\nabla \hat{f}_n(x)$ and $\nabla^2 \hat{f}_n(x)$ that satisfy

$$
(Z_n^T W_n Z_n) \left( \begin{array}{c} \nabla \hat{f}_n(x) \\ \nabla^2 \hat{f}_n(x) \end{array} \right) = Z_n^T W_n \hat{f}(x, Y_n, N_n) \tag{56}
$$

Noting that $Z_n = \left[ Y_n \quad Y_n^Q \right]$, the equation above can be written as

$$
Y_n^T W_n Y_n \nabla \hat{f}_n(x) + Y_n^T W_n Y_n^Q \left[ \nabla^2 \hat{f}_n(x) \right]_{v} = Y_n^T W_n \hat{f}(x, Y_n, N_n)
$$

$$(Y_n^Q)^T W_n Y_n \nabla \hat{f}_n(x) + (Y_n^Q)^T W_n Y_n^Q \left[ \nabla^2 \hat{f}_n(x) \right]_{v} = (Y_n^Q)^T W_n \hat{f}(x, Y_n, N_n)
$$

Suppose we know that even though $Z_n^T W_n Z_n$ may not be positive definite, $Y_n^T W_n Y_n$ is positive definite. Then, $Y_n^T W_n Y_n$ is non-singular. Hence, any solution of (56) also satisfies

$$
\nabla \hat{f}_n(x) = (Y_n^T W_n Y_n)^{-1} Y_n^T W_n \left\{ \hat{f}(x, Y_n, N_n) - Y_n^Q \left[ \nabla^2 \hat{f}_n(x) \right]_{v} \right\} \tag{57}
$$

With the notation we have developed so far, we are ready to state and prove our convergence results. Consider any open set $E \subset \mathbb{R}^l$, a compact set $D \subset E$, and a sequence $\{x_n\}_{n \in \mathbb{N}} \subset D$. For each $n \in \mathbb{N}$, let $\nabla \hat{f}_n(x_n)$ and $\nabla^2 \hat{f}_n(x_n)$ be chosen to satisfy

$$
\nabla \hat{f}_n(x_n) = (Y_n^T W_n Y_n)^{-1} Y_n^T W_n \left\{ \hat{f}(x_n, Y_n, N_n) - Y_n^Q \left[ \nabla^2 \hat{f}_n(x_n) \right]_{v} \right\} \tag{58}
$$

Next, we motivate and state in detail the assumptions needed to get that

$$
\lim_{n \to \infty} \left\| \nabla \hat{f}_n(x_n) - \nabla f(x_n) \right\|_2 = 0
$$

Let us start with our assumption regarding the sequence $\{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}}$ of approximating functions and the function $f$.

**A 4.3.** For any compact set $D \subset E$, we have that $f \in C_1(D)$, $\{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \subset W_0(D)$ and

$$
\lim_{N \to \infty} \left\| \hat{f}(\cdot, N) - f \right\|_{W_0(D)} = 0
$$

Next, we state our assumptions regarding the perturbation vectors $y^i_n$ for each $n \in \mathbb{N}$ and $i \in \{1, \ldots, M_n\}$. Note that a perturbation $y^i_n = 0$ does not add any value to the objective function in (38) or (45) and adds only a redundant equation of the form $0 = 0$ in (43) and (56). Therefore, we assume that none of the perturbations are trivial. Also, for the exact same reasons mentioned above, we will also assume that the weight associated with each design point, is also positive.
A 4.4. For each \( n \in \mathbb{N} \) and \( i \in \{1, \ldots, M_n\} \), \( \|y_n^i\|_2 > 0 \) and \( w_n^i > 0 \).

Recall that in the case of finite differences, the step-size sequence \( \{c_n\}_{n \in \mathbb{N}} \) had to converge to zero as \( n \to \infty \) in order to obtain convergence \( \nabla \hat{f}_n(x) \to \nabla f(x) \). It is intuitive to expect that in the same fashion, for any \( n \in \mathbb{N} \) the accuracy of \( \nabla \hat{f}_n(x_n) \) determined as in (58) depend on the Euclidean distances \( \{\|y_n^i\|_2 : i = 1, \ldots, M_n\} \) of the design points \( \{x_n + y_n^i : i = 1, \ldots, M_n\} \) from \( x_n \). Indeed, we should expect that in order to ensure \( \|\nabla \hat{f}_n(x_n) - \nabla f(x_n)\|_2 \to 0 \) as \( n \to \infty \), the Euclidean distances between the design points and the point \( x_n \) must decrease to zero as \( n \to \infty \).

Thus, in order to monitor and control the Euclidean distances of the design points from \( x_n \), we define a neighborhood of \( x_n \) called the design region for each \( n \), as follows.

\[
D_n := \left\{ x \in \mathbb{R}^l : \|x - x_n\|_2 \leq \delta_n \right\}
\] (59)

We will refer to \( \delta_n > 0 \) as the design region radius for iteration \( n \). Without loss of generality, we will assume that \( \|y_n^i\|_2 \leq \delta_n \) for \( i = 1, \ldots, M_n^I \) for some \( M_n^I \leq M_n \) and \( \|y_n^i\|_2 > \delta_n \) for \( i = M_n^I + 1, \ldots, M_n \). We use the terms inner and outer to denote the design points lying respectively within and outside the design region. In order to obtain the convergence of \( \nabla \hat{f}_n(x_n) \) to \( \nabla f(x_n) \), we will ensure that \( \delta_n \to 0 \) as \( n \to \infty \).

A 4.5. The sequence \( \{\delta_n\}_{n \in \mathbb{N}} \) of design region radii is such that \( \delta_n > 0 \) for each \( n \in \mathbb{N} \) and \( \delta_n \to 0 \) as \( n \to \infty \).

Assumption A 4.5 immediately leads to the following result.

**Lemma 4.3.** Consider any open set \( \mathcal{E} \subset \mathbb{R}^l \), a compact set \( \mathcal{D} \subset \mathcal{E} \), and a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D} \).

Let Assumption A 4.5 hold. Then there exists \( \delta_D > 0 \) and \( M \in \mathbb{N} \), such that \( [x_n, x_n + y_n^i] \subset B(x_n, \delta_D) \subset \mathcal{E} \) for all \( n > M \) and \( i \in \{1, \ldots, M_n^I\} \). Further, there exists a compact set \( \mathcal{D}^* \) such that \( \mathcal{D} \subset \mathcal{D}^* \subset \mathcal{E} \) and \( [x_n, x_n + y_n^i] \subset \mathcal{D}^* \) for all \( n > N \) and \( i \in \{1, \ldots, M_n^I\} \).

**Proof.** Since \( \mathcal{E} \) is open and \( \mathcal{D} \subset \mathcal{E} \) is compact, it follows from Lemma 2.1 that there exists \( \delta_D > 0 \) such that \( B(x, \delta_D) \subset \mathcal{E} \) for all \( x \in \mathcal{D} \). In particular, since \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D} \), \( B(x_n, \delta_D) \subset \mathcal{E} \) for each \( n \in \mathbb{N} \).

From Assumption A 4.5, we know that since \( \delta_n \to 0 \) as \( n \to \infty \), there exists \( M \in \mathbb{N} \) such that for all \( n > M \), \( \delta_n < \delta_D \). Using the definition of \( D_n \) in (59), we get that \( x_n + y_n^i \in B(x_n, \delta_D) \) for
all \( n > M \) and all \( i \in \{1, \ldots, M^I_n\} \). Because \( B(x_n, \delta_D) \) is convex, the line segment \([x_n, x_n + y^i_n]\) is contained in \( B(x_n, \delta_D) \) for all \( n > M \) and all \( i \in \{1, \ldots, M^I_n\} \).

Further, setting
\[
D^* := \bigcup_{x \in D} B(x, \delta_D)
\]
it follows that \([x_n, x_n + y^i_n] \subset D^* \) for all \( n > M \) and all \( i \in \{1, \ldots, M^I_n\} \). It follows from Lemma 2.1 that \( D^* \) is compact and \( D \subset D^* \subset \mathcal{E} \).

Thus, through Assumption A 4.5 and Lemma 4.3 we have ensured that all the inner design points lie in a compact set \( D^* \subset \mathcal{E} \) for all \( n > M \in \mathbb{N} \). However, we need to explicitly ensure that the outer design points lie in a compact set \( \mathcal{D}^* \).

**A 4.6.** There exists a compact set \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D} \subset \mathcal{E} \) such that \( x_n + y^i_n \in \mathcal{D} \) for all \( n \in \mathbb{N} \) and \( i \in \{M^I_n + 1, \ldots, M_n\} \).

Now, from Lemma 4.3 and Assumption A 4.6 we get the following proposition.

**Proposition 4.4.** Let \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D} \) where \( \mathcal{D} \subset \mathcal{E} \) is compact. Let Assumption A 4.5 and Assumption A 4.6 hold. Then there exists a compact set \( \mathcal{D}^* \subset \mathcal{E} \) containing \( \mathcal{D} \) and \( M \in \mathbb{N} \) such that \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D}^* \) and for all \( n > M \) and \( i \in \{1, \ldots, M_n\} \), \( x_n + y^i_n \in \mathcal{D}^* \).

For the notation that we defined earlier related to the design points, we will use the superscripts “\( I \)” and “\( O \)” to denote the corresponding notation for the inner and outer design points respectively. Thus, we let

\[
Y^I_n := \begin{pmatrix} (y^I_n)^T \\ \vdots \\ (y^{M^I_n}_n)^T \end{pmatrix} \quad \text{and} \quad Y^O_n := \begin{pmatrix} (y^{M^I_n+1}_n)^T \\ \vdots \\ (y^{M_n}_n)^T \end{pmatrix}
\]  

(60)

Similarly, let

\[
[Y^I_n]^Q := \begin{pmatrix} ([y^I_n]^Q)^T \\ \vdots \\ ([y^{M^I_n}]^Q)^T \end{pmatrix}, \quad [Y^O_n]^Q := \begin{pmatrix} ([y^{M^I_n+1}_n]^Q)^T \\ \vdots \\ ([y^{M_n}_n]^Q)^T \end{pmatrix}
\]  

(61)

and

\[
Z^I_n := \begin{pmatrix} Y^I_n \quad [Y^I_n]^Q \end{pmatrix}, \quad Z^O_n := \begin{pmatrix} Y^O_n \quad [Y^O_n]^Q \end{pmatrix}
\]  

(62)
Thus,
\[
Y_n = \begin{pmatrix} Y^I_n \\ Y^O_n \end{pmatrix} \quad \text{and} \quad Z_n = \begin{pmatrix} Z^I_n \\ Z^O_n \end{pmatrix}
\]

Analogous to \( \mathcal{N}_n \), we let \( \mathcal{N}^I_n := \{N^1_n, \ldots, N^{M^I_n}_n\} \) and \( \mathcal{N}^O_n := \{N^{M^I_n+1}_n, \ldots, N^{M_n}_n\} \) denote the set of sample sizes used to evaluate the sample average functions at the inner and outer design points respectively. Then, the vectors \( \hat{f}(x, Y^I_n, \mathcal{N}^I_n) \subset \mathbb{R}^{M^I_n} \) and \( \hat{f}(x, Y^O_n, \mathcal{N}^O_n) \subset \mathbb{R}^{M_n-M^I_n} \) are defined analogous to \( \hat{f}(x, Y_n, \mathcal{N}_n) \). Also, we let \( W^I_n = \text{diag}(w^1_n, \ldots, w^{M^I_n}_n) \) and \( W^O_n = \text{diag}(w^{M^I_n+1}_n, \ldots, w^{M_n}_n) \).

Ensuring \( \delta_n \to 0 \) alone is insufficient to ensure the convergence of \( \| \nabla \hat{f}_n(x_n) - \nabla f(x) \|_2 \) to zero. To see this, recall that in our finite different example, all our design points converged to \( x \) as \( n \to \infty \). However, in our formulation we allow for so-called inner and outer design points and \( \delta_n \to 0 \) only ensure only that the the inner design points converge toward the point \( x_n \) as \( n \to \infty \). Further, we have placed no restrictions on positions of the inner design points within the design region. Accordingly, We need to enforce two more conditions on the placement of the inner design points in relation to the point \( x_n \) and the weights associated with both the inner and outer design points in order to ensure that \( \| \nabla \hat{f}_n(x_n) - \nabla f(x_n) \|_2 \to 0 \) as \( n \to \infty \).

**A 4.7.** The following conditions hold:

1. There exists \( K^I_\lambda < \infty \) such that
   \[
   \left\| (Y^I_n)^T W^I_n Y^I_n \right\|_2 \left\| (Y^I_n)^T W^I_n Y^I_n \right\|_2 < K^I_\lambda \quad \text{for all} \quad n \in \mathbb{N} \quad (63)
   \]

2. \[
   \lim_{n \to \infty} \frac{\| (Y^O_n)^T W^O_n Y^O_n \|_2}{\| (Y^I_n)^T W^I_n Y^I_n \|_2} = 0 \quad (64)
   \]

The first condition in Assumption A 4.7 places a limit on the condition number of the sequence \( \{(Y^I_n)^T W^I_n Y^I_n\}_{n \in \mathbb{N}} \). This ensures that the perturbations are chosen so as to avoid the problem of multicollinearity, which occurs when all the perturbation vectors lie close to a subspace of dimension less than \( l \). Note that this condition automatically requires that there must exist at least \( l \) inner design points for each \( n \in \mathbb{N} \). It is easy to choose a sequence of perturbations \( \{y^1_n, \ldots, y^{M_n}_n\}_{n \in \mathbb{N}} \) such that the first condition in Assumption A 4.7 is satisfied. For example, we can choose \( M \geq l \).
vectors $y^1, \ldots, y^M \in \mathbb{R}^l$ that span $\mathbb{R}^l$, and let $Y \in \mathbb{R}^{M \times l}$ be given by

$$Y := \begin{pmatrix} (y^1)^T \\ \vdots \\ (y^M)^T \end{pmatrix}$$

One simple example of such a choice is $M = l$ and $y^i = e_i$, where $\{e_1, \ldots, e_l\}$ is the standard basis for $\mathbb{R}^l$, as in the case of finite differences. It follows that $Y^TY \in \mathbb{S}_{l \times l}$ is positive definite. Then, for some sequence $\{\delta_n\}_{n \in \mathbb{N}}$ such that $\delta_n \to 0$ as $n \to \infty$, we can choose $M^I_n = M$, $y^i_n = \delta_n y^i$ and $w^i_n = 1$ for each $n \in \mathbb{N}$ and each $i \in \{1, \ldots, M\}$. It follows that the condition number of $(Y^I_n)^TW_n Y_n$ is equal to the condition number of $Y^TY$ for each $n \in \mathbb{N}$ and this bounded.

The second condition in Assumption A 4.7 ensures that the outer design points exert a progressively decreasing influence on $\nabla \hat{f}_n(x_n)$ as $n \to \infty$. It is easy to choose weights $\{w^i_n : i = 1, \ldots, M_n\}$ such that the second condition in Assumption A 4.7 is satisfied. Let

$$\pi_n := \sum_{i=1}^{M^I_n} \|y^i_n\|^2_2$$

and let the weights $w^i_n$ be set as follows:

$$w^i_n = \begin{cases} 1 & \text{for } i \in \{1, \ldots, M^I_n\} \\ \frac{\delta_n \pi_n}{(M_n-M^I_n)\|y^i_n\|^2_2} & \text{for } i \in \{M^I_n+1, \ldots, M_n\} \end{cases}$$

(65)

It follows from (7) that

$$\frac{\| (Y^O_n)^TW_n^O Y_n^O \|_2}{\| (Y^I_n)^TW_n Y_n \|_2} \leq l \frac{\text{trace}((Y^O_n)^TW_n^O Y_n^O)}{\text{trace}((Y^I_n)^TW_n Y_n)}$$

Note that

$$\text{trace}(Y^TWY) = \sum_{j=1}^l \sum_{i=1}^m w^i Y^2_{ij}$$

$$= \sum_{i=1}^m w^i \sum_{j=1}^l Y^2_{ij}$$

$$= \sum_{i=1}^m w^i \|y^i\|^2_2$$

36
Therefore,

\[
\frac{\| (Y_n^O)^T W_n^O Y_n^O \|}{\| (Y_n^I)^T W_n^I Y_n^I \|} \leq \frac{l \sum_{i=M_n^I+1}^{M_n^I} w_n^i \| y_n^I \|^2}{\sum_{i=1}^{M_n^I} w_n^i \| y_n^I \|^2} = \frac{l \delta_n \pi_n}{\sum_{i=1}^{M_n^I} \| y_n^I \|^2} = l \delta_n
\]

The first equality above follows from the substitution of the example weights in (65). Therefore, since \( \delta_n \to 0 \) as \( n \to \infty \), it follows that

\[
\lim_{n \to \infty} \frac{\| (Y_n^O)^T W_n^O Y_n^O \|}{\| (Y_n^I)^T W_n^I Y_n^I \|} = 0
\]

Next, we make an assumption regarding the set of approximation parameters \( \{N_n^1, \ldots, N_n^{M_n^I}\} \)

**A 4.8.** The sample sizes \( N_n^i \) corresponding to the inner design points all increase to infinity as \( n \to \infty \). That is,

\[
\lim_{n \to \infty} \min_{i=1, \ldots, M_n^I} N_n^i = \infty \quad (66)
\]

If \( f \) is an expected value function as in (11) and \( \hat{f}(\cdot, N)(\omega) \) is the sample average function, then we allow the sample sizes \( \{N_n^i : i = 1, \ldots, M_n^I\} \) to be random variables and allow for (66) to hold in the \( P \)-almost sure sense.

Finally, we make an assumption regarding our sequence of Hessian approximations \( \{\nabla^2 \hat{f}_n(x_n)\}_{n \in \mathbb{N}} \).

**A 4.9.** The sequence of Hessian approximation matrices \( \{\nabla^2 \hat{f}_n(x_n)\}_{n \in \mathbb{N}} \) is norm-bounded. That is, there exists \( K_H < \infty \) such that \( \| \nabla^2 \hat{f}_n(x_n) \|_2 < K_H \) for all \( n \in \mathbb{N} \).

We will first show a couple of useful Lemmas and proceed to show that \( \| \nabla \hat{f}_n(x_n) - \nabla f(x_n) \|_2 \to 0 \) as \( n \to \infty \).

**Lemma 4.5.** Suppose that Assumption A 4.7 holds. Then the following hold:

1.

\[
\| (Y_n^T W_n Y_n)^{-1} \|_2 \| (Y_n^I)^T W_n^I Y_n^I \|_2 < K_n^I \quad \text{for all } n \in \mathbb{N} \quad (67)
\]

2.

\[
\lim_{n \to \infty} \| (Y_n^T W_n Y_n)^{-1} \|_2 \| (Y_n^O)^T W_n^O Y_n^O \|_2 = 0 \quad (68)
\]
3. There is a constant $K_\lambda < \infty$ such that

$$\| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n Y_n \|_2 < K_\lambda \quad \text{for all } n \in \mathbb{N} \quad (69)$$

Proof. Note that

$$Y_n^T W_n Y_n = (Y_n^I)^T W_n^I Y_n^I + (Y_n^O)^T W_n^O Y_n^O$$

Since $(Y_n^O)^T W_n^O Y_n^O$ is positive semidefinite, it follows from the interleaving eigenvalues theorem that

$$\lambda^{\text{min}}(Y_n^T W_n Y_n) \geq \lambda^{\text{min}}((Y_n^I)^T W_n^I Y_n^I) > 0$$

Thus,

$$\| (Y_n^T W_n Y_n)^{-1} \|_2 \leq \frac{1}{\lambda^{\text{min}}(Y_n^I)^T W_n Y_n^I) \leq \frac{1}{\lambda^{\text{min}}(Y_n^I)^T W_n Y_n^I) \leq \| (Y_n^I)^T W_n Y_n^I \|_2$$

Therefore, it follows from (63) and (70) that

$$\| (Y_n^T W_n Y_n)^{-1} \|_2 \| (Y_n^I)^T W_n^I Y_n^I \|_2 \leq \| (Y_n^I)^T W_n^I Y_n^I \|_2 \| (Y_n^I)^T W_n^I Y_n^I \|_2 < K_\lambda$$

Thus, we have shown that (67) holds. Next,

$$\| (Y_n^T W_n Y_n)^{-1} \|_2 \| (Y_n^I)^T W_n^O Y_n^O \|_2 = \| (Y_n^I)^T W_n^I Y_n^I \|_2 \| (Y_n^O)^T W_n Y_n^O \|_2 \| (Y_n^I)^T W_n^I Y_n^I \|_2$$

It follows from (67) that

$$\| (Y_n^T W_n Y_n)^{-1} \|_2 \| (Y_n^I)^T W_n^O Y_n^O \|_2 \leq K_\lambda \| (Y_n^O)^T W_n Y_n^O \|_2 \| (Y_n^I)^T W_n^I Y_n^I \|_2$$

It follows from (64) in Assumption A 4.7 that

$$\lim_{n \to \infty} \| (Y_n^T W_n Y_n)^{-1} \|_2 \| (Y_n^O)^T W_n^O Y_n^O \|_2 = 0$$

Next, (69) follows by noting that

$$\| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n Y_n \|_2 = \| (Y_n^T W_n Y_n)^{-1} \|_2 \| (Y_n^I)^T W_n^I Y_n^I \|_2 + (Y_n^O)^T W_n^O Y_n^O \|_2$$

$$\leq \| (Y_n^T W_n Y_n)^{-1} \|_2 \| (Y_n^I)^T W_n^I Y_n^I \|_2 + \| (Y_n^T W_n Y_n)^{-1} \|_2 \| (Y_n^O)^T W_n^O Y_n^O \|_2$$

and using (67) and (68).
Lemma 4.6. Consider any matrix $Y \in \mathbb{R}^{m \times l}$ and a positive diagonal matrix $W \in \mathbb{R}^{m \times m}$. Let $y_i \in \mathbb{R}^l$ denote row $i$ of $Y$, and let $w_i \in \mathbb{R}$ denote diagonal element $W_{ii}$, $i = 1, \ldots, m$. For some $m_1 \leq m$, let

$$Y^I = \begin{pmatrix} y_1 \\ \vdots \\ y_{m_1} \end{pmatrix} \quad \text{and} \quad Y^O = \begin{pmatrix} y_{m_1+1} \\ \vdots \\ y_m \end{pmatrix}$$

Similarly, let $W^I = \text{diag}(w_1, \ldots, w_{m_1})$ and $W^O = \text{diag}(w_{m_1+1}, \ldots, w_m)$.

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ be a set of real numbers and $\{v^1, v^2, \ldots, v^m\}$ be a set of vectors in $\mathbb{R}^l$. Also, let $a, b, c \in \mathbb{R}^m$ be given by

$$a := \begin{pmatrix} \|y_1\|_2 \alpha_1 \\ \|y_2\|_2 \alpha_2 \\ \vdots \\ \|y_m\|_2 \alpha_m \end{pmatrix} \quad b := \begin{pmatrix} y_1^T v^1 \\ y_2^T v^2 \\ \vdots \\ y_m^T v^m \end{pmatrix} \quad \text{and} \quad c := \begin{pmatrix} y_1^T v^1 \\ \vdots \\ y_{m_1}^T v^m \\ \|y_{m_1+1}\|_2 \alpha_{m_1+1} \\ \vdots \\ \|y_m\|_2 \alpha_m \end{pmatrix}$$

Then,

$$\|Y^T W a\|_2 \leq l \left[ \|Y^I Y^I\| \max_{2 \in \{1, \ldots, m_1\}} \|\alpha_i\| + \|Y^O Y^O\| \max_{2 \in \{m_1+1, \ldots, m\}} \|\alpha_i\| \right]$$

(71)

$$\|Y^T W b\|_2 \leq l \left[ \|Y^I Y^I\| \max_{2 \in \{1, \ldots, m_1\}} \|v^i\|_2 + \|Y^O Y^O\| \max_{2 \in \{m_1+1, \ldots, m\}} \|v^i\|_2 \right]$$

(72)

$$\|Y^T W c\|_2 \leq l \left[ \|Y^I Y^I\| \max_{2 \in \{1, \ldots, m_1\}} \|v^i\|_2 + \|Y^O Y^O\| \max_{2 \in \{m_1+1, \ldots, m\}} \|\alpha_i\| \right]$$

(73)
Proof. We first prove (71).

\[ \|Y^T W a\|_2 \leq \sum_{i=1}^{m} y_i w_i \|y_i\|_2 \alpha_i \| \leq \sum_{i=1}^{m} \|y_i w_i\|_2 \alpha_i \| \]

\[ = \sum_{i=1}^{m} w_i \|y_i\|_2 |\alpha_i| \]

\[ = \sum_{i=1}^{m_1} w_i \|y_i\|_2 |\alpha_i| + \sum_{i=m_1+1}^{m} w_i \|y_i\|_2 |\alpha_i| \]

\[ \leq \sum_{i=1}^{m_1} w_i \|y_i\|_2 |\alpha_k| + \sum_{i=m_1+1}^{m} w_i \|y_i\|_2 |\alpha_k| \]

\[ = \sum_{i=1}^{m_1} \sum_{j=1}^{l} w_i y_{ij}^2 \max_{k \in \{1, ..., m_1\}} |\alpha_k| + \sum_{i=m_1+1}^{m} \sum_{j=1}^{l} w_i y_{ij}^2 \max_{k \in \{m_1+1, ..., m\}} |\alpha_k| \]

\[ = \text{trace}(Y^T W Y^T) \max_{k \in \{1, ..., m_1\}} |\alpha_k| + \text{trace}(Y^T W Y^T) \max_{k \in \{m_1+1, ..., m\}} |\alpha_k| \]

\[ \leq l \left[ \left\| Y^T W Y \right\|_{2 \max_{i \in \{1, ..., m_1\}} |\alpha_i|} + \left\| Y^T W Y \right\|_{2 \max_{i \in \{m_1+1, ..., m\}} |\alpha_i|} \right] \]

Equation (72) can be established in a similar fashion, by using the Cauchy-Schwartz inequality

\[ |(y_i)^T v| \leq \|y_i\| \|v\| \]

\[ \|Y^T W b\|_2 \leq \sum_{i=1}^{m} \|y_i w_i (y_i)^T v\|_2 \]

\[ = \sum_{i=1}^{m} w_i |(y_i)^T v| \|y_i\|_2 \]

\[ \leq \sum_{i=1}^{m_1} w_i \|y_i\|_2 \|v\|_2 \]

\[ \leq \sum_{i=1}^{m_1} w_i \|y_i\|_2 \|v\|_2 + \sum_{i=m_1+1}^{m} w_i \|y_i\|_2 \|v\|_2 \]

\[ \leq l \left[ \left\| Y^T W Y \right\|_{2 \max_{i \in \{1, ..., m_1\}} \|v\|_2} + \left\| Y^T W Y \right\|_{2 \max_{i \in \{m_1+1, ..., m\}} \|v\|_2} \right] \]
Similarly, we get
\[
\|Y^TWc\|_2 = \left\| \sum_{i=1}^{m_1} y_iw_i(y_i)^T v^i + \sum_{i=m_1+1}^{m} y_iw_i \right\|_2
\leq \sum_{i=1}^{m_1} \|y_iw_i(y_i)^T v^i\|_2 + \sum_{i=m_1+1}^{m} \|y_iw_i\|_2 \alpha_i
\leq \sum_{i=1}^{m_1} w_i \|y_i\|_2^2 \|v^i\|_2 + \sum_{i=m_1+1}^{m} w_i \|y_i\|_2^2 |\alpha^i|
\leq l \left[ \max_{2i \in \{1, \ldots, m_1\}} \|v^i\|_2^2 + \max_{2i \in \{m_1+1, \ldots, m\}} |\alpha^i| \right]
\]

\[ \square \]

**Theorem 4.7.** Consider a sequence \(\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D} \subset \mathcal{E}\) where \(\mathcal{D}\) is compact and let for each \(n \in \mathbb{N}\), \(\hat{\nabla}_n f(x_n) \in \mathbb{R}^l\) and \(\hat{\nabla}_n^2 f(x_n) \in \mathbb{S}^{l \times l}\) be picked so as to satisfy (58). Suppose that the Assumptions A 4.3 through A 4.9 hold. Then, it holds that
\[
\lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 = 0
\]
In particular, if \(\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D}\) is such that \(x_n \to \bar{x} \in \mathcal{D}\) as \(n \to \infty\), then
\[
\lim_{n \to \infty} \hat{\nabla}_n f(x_n) = \nabla f(\bar{x})
\]

**Proof of Theorem 4.7:**

From Lemma 4.5, we know that \(Y_n^TW_nY_n\) is non-singular for each \(n \in \mathbb{N}\). Therefore, we can rewrite (58) for each \(n \in \mathbb{N}\), as
\[
\nabla \hat{f}_n(x_n) = (Y_n^TW_nY_n)^{-1}Y_n^TW_n\hat{f}(x_n, Y_n, N_n) - (Y_n^TW_nY_n)^{-1}(Y_n^TW_nY_n^Q) \left[ \hat{\nabla}_n^2 f_n(x_n) \right]_v \tag{74}
\]
We proceed by manipulating the above expression for \(\nabla \hat{f}_n(x_n)\) given and showing that \(\left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 \to 0\) as \(n \to \infty\). From (74), we get
\[
\nabla \hat{f}_n(x_n) - \nabla f(x_n) = (Y_n^TW_nY_n)^{-1}Y_n^TW_n \left\{ \hat{f}(x_n, Y_n, N_n) - Y_n^Q \left[ \hat{\nabla}_n^2 f_n(x_n) \right]_v \right\} - (Y_n^TW_nY_n)^{-1}Y_n^TW_nY_n \nabla f(x_n)
\]
Adding and subtracting \((Y_n^TW_nY_n)^{-1}Y_n^TW_nf(x_n, Y_n)\), we get
\[
\nabla \hat{f}_n(x_n) - \nabla f(x_n) = (Y_n^TW_nY_n)^{-1}Y_n^TW_n \left[ \hat{f}(x_n, Y_n, N_n) - f(x_n, Y_n) \right] +
(Y_n^TW_nY_n)^{-1}Y_n^TW_n \left[ f(x_n, Y_n) - Y_n \nabla f(x_n) \right] - (Y_n^TW_nY_n)^{-1}Y_n^TW_nY_n^Q \left[ \hat{\nabla}_n^2 f_n(x_n) \right]
\]

41
From Assumption A 4.4 we know that \( \|y_i^k\|_2 > 0 \) for \( i = 1, \ldots, M_n \). Using this we define

\[
\begin{align*}
a_n & := \left( \hat{f}(x_n, Y_n, N_n) - f(x_n, Y_n) \right) \\
& = \left( \|y_n^k\|_2 \frac{(f(x_n + y_n^k, N_n) - f(x_n, N_n)) - (f(x_n + y_n^k) - f(x_n))}{\|y_n^k\|_2} \right) \\
& \quad \vdots \\
& = \left( \|y_n^{M_n}\|_2 \frac{(f(x_n + y_n^{M_n}, N_n^{M_n}) - f(x_n, N_n^{M_n})) - (f(x_n + y_n^{M_n}) - f(x_n))}{\|y_n^{M_n}\|_2} \right)
\end{align*}
\] (75)

Also, let

\[
\begin{align*}
b_n & := f(x_n, Y_n) - Y_n \nabla f(x_n) \\
& = \left( f(x_n + y_n^1) - f(x_n) - y_n^1 \nabla f(x_n) \right) \\
& \quad \vdots \\
& = \left( f(x_n + y_n^{M_n}) - f(x_n) - y_n^{M_n} \nabla f(x_n) \right) = \left( \left\|y_n^k\right\|_2 \frac{f(x_n + y_n^k) - f(x_n) - y_n^k \nabla f(x_n)}{\left\|y_n^k\right\|_2} \right) \\
& \quad \vdots \\
& = \left( \left\|y_n^{M_n}\right\|_2 \frac{f(x_n + y_n^{M_n}) - f(x_n) - y_n^{M_n} \nabla f(x_n)}{\left\|y_n^{M_n}\right\|_2} \right)
\end{align*}
\] (76)

Finally, we set

\[
\begin{align*}
c_n & = Y_n^{Q} \left[ \hat{\nabla}_n^2 f(x_n) \right]_v \\
& = \left( \left( y_n^1 \right)^T \left[ \hat{\nabla}_n^2 f(x_n) \right]_v \right) \\
& \quad \vdots \\
& = \left( \left( y_n^{M_n} \right)^T \left[ \hat{\nabla}_n^2 f(x_n) \right]_v \right)
\end{align*}
\] (77)

(78)

Using the notation in (49) and the fact that \( \|y_i^k\|_2 > 0 \) for all \( i = 1, \ldots, M_n \) and \( n \in \mathbb{N} \), we get

\[
\begin{align*}
c_n & = \left( \frac{1}{2} y_n^1 \hat{\nabla}_n^2 f(x_n) y_n^1 \right) \\
& \quad \vdots \\
& = \left( \frac{1}{2} y_n^{M_n} \hat{\nabla}_n^2 f(x_n) y_n^{M_n} \right)
\end{align*}
\] (78)

Then, using the notation defined above, we get

\[
\hat{\nabla}_n f(x_n) - \nabla f(x) = (Y_n^{T} W_n Y_n)^{-1} Y_n^{T} W_n a_n + (Y_n^{T} W_n Y_n)^{-1} Y_n^{T} W_n b_n - (Y_n^{T} W_n Y_n)^{-1} Y_n^{T} W_n c_n
\]

Using the triangle inequality,

\[
\| \hat{\nabla}_n f(x_n) - \nabla f(x) \|_2 \leq \| (Y_n^{T} W_n Y_n)^{-1} Y_n^{T} W_n a_n \|_2 + \| (Y_n^{T} W_n Y_n)^{-1} Y_n^{T} W_n b_n \|_2 + \| (Y_n^{T} W_n Y_n)^{-1} Y_n^{T} W_n c_n \|_2
\]

(79)

Let us consider the three terms on the right hand side of (79) in order.
Using (71) in Lemma 4.6, we get
\[
\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n a_n \|_2 \leq \| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n a_n \|_2
\]
\[
\leq l \| (Y_n^T W_n Y_n)^{-1} \|_2 \left\{ \max_{i \in \{1, \ldots, M_n^i \}} \left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - (f(x_n + y_n^i) - f(x_n)) \right| \right\}
\]
\[
+ \left\{ \| Y_n^O^T W_n^O Y_n^O \|_2 \max_{i \in \{M_n^i + 1, \ldots, M_n \}} \left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - (f(x_n + y_n^i) - f(x_n)) \right| \right\}
\]
(80)

Now, we consider the two terms of (80) in order. First, from (67) in Lemma 4.5, we know that for all \( n \in \mathbb{N}, (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n Y_n \|_2 \leq K^\lambda_\nu. \) From Assumption A 4.8, we know that
\[
\lim_{n \to \infty} \min_{i = 1, \ldots, M_n^i} N_n^i = \infty
\]
Also, we get from Proposition P 4.4, we know that \( \{x_n\}_{n \in \mathbb{N}} \in \mathcal{D} \subset \hat{\mathcal{D}} \) and for all \( n > M, x_n + y_n^i \in \hat{\mathcal{D}} \) for each \( i = 1, \ldots, M_n. \) Now, from Assumption A 4.3 and the definition of the norm on \( \mathbb{R}_0(\hat{\mathcal{D}}), \) we get for any \( \varepsilon > 0 \) there exists \( N_\varepsilon \in \mathbb{N} \) such that for all \( N > N_\varepsilon, \)
\[
\sup_{x+y, x \in \mathcal{D}} \sup_{y \neq 0} \left| \left( \hat{f}(x+y, N) - \hat{f}(x, N) \right) - (f(x+y) - f(x)) \right| \leq \varepsilon
\]
Also, from Assumption A 4.8, we know that there exists \( \tilde{N} \in \mathbb{N} \) such that for all \( n > \tilde{N}, \) \( \min\{N_n^i : i = 1, \ldots, M_n^i \} > N_\varepsilon. \) Thus, we get combining these two facts, that for any \( \varepsilon > 0, \) there exists \( \tilde{N} > M, \) such that for all \( n > \tilde{N}, \)
\[
\max_{i \in \{1, \ldots, M_n^i \}} \left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - (f(x_n + y_n^i) - f(x_n)) \right| \leq \varepsilon
\]
Therefore,
\[
\lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^i \}} \left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - (f(x_n + y_n^i) - f(x_n)) \right| = 0
\]
Consequently, we get that,
\[
\lim_{n \to \infty} \| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n Y_n \|_2 \left\{ \max_{i \in \{1, \ldots, M_n^i \}} \left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - (f(x_n + y_n^i) - f(x_n)) \right| \right\} = 0
\]

Next consider the second term on the right side of (80). First, from Lemma 4.5, we get
\[
\lim_{n \to \infty} \| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^O^T W_n^O Y_n^O \|_2 = 0
\]
43
Thus, we have shown that

\[ \max_{i \in \{M'_n + 1, \ldots, M_n\}} \left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - \left( f(x_n + y_n^i) - f(x_n) \right) \right| \frac{\|y_n^i\|_2}{\|y_n^i\|_2} \leq \max_{i \in \{M'_n + 1, \ldots, M_n\}} \left\{ \frac{\|\hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i)\|}{\|y_n^i\|_2} + \frac{\|\hat{f}(x_n + y_n^i) - f(x_n)\|}{\|y_n^i\|_2} \right\} \leq \max_{i \in \{M'_n + 1, \ldots, M_n\}} \left\{ \frac{\sup_{x, x+y \in \tilde{D}, y \neq 0} |\hat{f}(x+y, N_n^i) - \hat{f}(x, N_n^i)|}{\|y\|_2} + \sup_{x, x+y \in \tilde{D}, y \neq 0} |f(x+y) - f(x)| \right\} \leq \|f\|_{W_0(\tilde{D})} + \max_{i \in \{M'_n + 1, \ldots, M_n\}} \left\| \hat{f}(\cdot, N_n^i) \right\|_{W_0(\tilde{D})} \]

Hence, the sequence \( \left\{ \|\hat{f}(\cdot, N)\|_{W_0(\tilde{D})} \right\}_{N \in \mathbb{N}} \) is bounded and there exists \( K_f < \infty \), such that \( \|f\|_{W_0(\tilde{D})} < K_f \) and \( \sup_{N \in \mathbb{N}} \left\| \hat{f}(\cdot, N)\|_{W_0(\tilde{D})} \right\| < K_f \) where \( \tilde{D} \subset \mathcal{E} \) is the compact set mentioned in Proposition P 4.4.

Consequently, we get that

\[ \|f\|_{W_0(\tilde{D})} + \max_{i \in \{M'_n + 1, \ldots, M_n\}} \left\| \hat{f}(\cdot, N_n^i) \right\|_{W_0(\tilde{D})} \leq 2K_f \]

Hence, using the fact that \( \lim_{n \to \infty} \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\| Y_n^{O^T} W_n^{O^T} Y_n^{O} \right\|_2 = 0 \), we finally get

\[ \lim_{n \to \infty} \left[ \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\| Y_n^{O^T} W_n^{O^T} Y_n^{O} \right\|_2 \times \max_{i \in \{M'_n + 1, \ldots, M_n\}} \left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - \left( f(x_n + y_n^i) - f(x_n) \right) \right| \right] = 0 \]

Thus, we have shown that

\[ \lim_{n \to \infty} \left\| (Y_n^T Y_n)^{-1} Y_n^T a_n \right\|_2 = 0 \]
Next, we consider the second term on the right side of (79). Using (71) in Lemma 4.6, we get

\[
\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n b_n \|_2 \leq l \| (Y_n^T W_n Y_n)^{-1} \|_2 \times \\
\left\{ \left\| Y_n^T W_n Y_n \right\|_{2 \to 1} \max_{i \in \{1, \ldots, M_n^I \}} \left| f(x_n + y_n^i) - f(x_n) - y_n^T T \nabla f(x_n) \right| \right\} \\
+ \left\{ \left\| Y_n^O W_n Y_n \right\|_{2 \to 1} \max_{i \in \{M_n^I + 1, \ldots, M_n \}} \left| f(x_n + y_n^i) - f(x_n) - y_n^T T \nabla f(x_n) \right| \right\} \tag{81}
\]

We will first show that

\[
\lim_{n \to \infty} l \| (Y_n^T W_n Y_n)^{-1} \|_2 \left\| Y_n^T W_n Y_n \right\|_{2 \to 1} \max_{i \in \{1, \ldots, M_n^I \}} \left| f(x_n + y_n^i) - f(x_n) - y_n^T T \nabla f(x_n) \right| = 0
\]

From Assumption A 4.3, we know that \( \nabla f : \tilde{D} \to \mathbb{R} \) is continuous on \( \tilde{D} \). Since the set \( \tilde{D} \subset \mathcal{E} \) (defined in Proposition P 4.4) is compact, we get that \( \nabla f \) is uniformly continuous on \( \tilde{D} \). That is, for any \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[
\sup_{x, x+y \in \tilde{D}} \sup_{0 < \| y \|_2 \leq \delta} \left| \frac{f(x+y) - f(x) - \nabla f(x) y}{\| y \|_2} \right| < \varepsilon \tag{82}
\]

From Proposition P 4.4, we know that \( \{x_n \}_{n \in \mathbb{N}} \subset \tilde{D} \) and \( \{x_n + y_n : i = 1, \ldots, M_n \} \subset \tilde{D} \) for \( n > M \). Further, we know from Assumption A 4.5 that \( \delta_n \to 0 \) as \( n \to \infty \). Since \( \| y_n^i \|_2 \leq \delta_n \) and \( 0 \leq t_n^i \leq 1 \) for each \( i = 1, \ldots, M_n^I \) and each \( n \in \mathbb{N} \), we get that

\[
\lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^I \}} \| x_n + y_n^i - x_n \|_2 = \lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^I \}} \| y_n^i \|_2 = 0 \tag{83}
\]

It follows from (83) that given any \( \delta > 0 \), there exists an \( M_1 \in \mathbb{N} \) with \( M_1 > M \) such that \( \| x_n + y_n^i - x_n \|_2 < \delta \) for all \( i \in \{1, \ldots, M_1^I \} \) and \( n > M_1 \). Therefore, using (82), we get that for any \( \varepsilon > 0 \), there exists \( M_1 \in \mathbb{N} \), such that

\[
\left| \frac{f(x_n + y_n^i) - f(x_n) - y_n^T T \nabla f(x_n)}{\| y_n^i \|_2} \right| < \varepsilon \quad \text{for all} \quad i = 1, \ldots, M_n^I \quad \text{and} \quad n > n_1
\]

Thus

\[
\lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^I \}} \left| \frac{f(x_n + y_n^i) - f(x_n) - y_n^T T \nabla f(x_n)}{\| y_n^i \|_2} \right| = 0
\]

Now, using (67) in Lemma 4.5, we get

\[
\lim_{n \to \infty} l \left( \| (Y_n^T W_n Y_n)^{-1} \|_2 \left\| Y_n^T W_n Y_n \right\|_{2 \to 1} \max_{i \in \{1, \ldots, M_n^I \}} \left| f(x_n + y_n^i) - f(x_n) - y_n^T T \nabla f(x_n) \right| \right) \\
\leq l K_n^I \lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^I \}} \left| \frac{f(x_n + y_n^i) - f(x_n) - y_n^T T \nabla f(x_n)}{\| y_n^i \|_2} \right| = 0
\]

45
Thus, the second term on the right side of (81) converges to zero. Hence we know from Proposition P 4.4 that
\[ \lim_{n \to \infty} \frac{\| (Y_n^T W_n Y_n)^{-1} \|_2}{\max_{2 \in \{L+1, \ldots, M_n\}} f(x_n + y_n^i) - f(x_n) - y_n^i \nabla f(x_n)} = 0 \]
Again, since \( f \) is continuously differentiable on \( \tilde{D} \) and \( \tilde{D} \subset \mathcal{E} \) is compact, we get from Lemma 2.3 that there exists \( K_{1f} < \infty \) such that
\[ \sup_{x + y, x \in \tilde{D}, y \neq 0} \frac{|f(x + y) - f(x) - y^T \nabla f(x)|}{\|y\|_2} < K_{1f} \]
Since we know from Proposition P 4.4 that \( x_n \in \tilde{D} \) for each \( n \in \mathbb{N} \) and \( x_n + y_n^i \in \tilde{D} \) for each \( n > M \) and \( i \in \{1, \ldots, M_n\} \), we get that for \( n > M \)
\[ \max_{i \in \{L+1, \ldots, M_n\}} f(x_n + y_n^i) - f(x_n) - y_n^i \nabla f(x_n) \leq K_{1f} \]
Now, using (68) in Lemma 4.5, we get
\[ \lim_{n \to \infty} l \| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n^O Y_n^O \|_2 \max_{2 \in \{L+1, \ldots, M_n\}} \frac{f(x_n + y_n^i) - f(x_n) - y_n^i \nabla f(x_n)}{\|y_n^i\|_2} \leq 2l \ K_{1f} \lim_{n \to \infty} \| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n^O Y_n^O \|_2 = 0 \]
Thus, the second term on the right side of (81) converges to zero. Hence \( \| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n b_n \|_2 \to 0 \) as \( n \to \infty \).

Finally, consider the third term on the right in (79). Again using (71),
\[ \| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n c_n \|_2 \leq \frac{l}{2} \| (Y_n^T W_n Y_n)^{-1} \|_2 \left\{ \| Y_n^T W_n^T W_n Y_n^O \|_2 \max_{i \in \{1, \ldots, M_n\}} \frac{y_n^i \nabla^2 \hat{f}_n(x_n) y_n^i}{\|y_n^i\|_2} \right\} \]
\[ + \left\{ \| Y_n^T W_n^O Y_n^O \|_2 \max_{i \in \{L+1, \ldots, M_n\}} \frac{y_n^i \nabla^2 \hat{f}_n(x_n) y_n^i}{\|y_n^i\|_2} \right\} \]
Since \( \nabla^2 \hat{f}_n(x_n) \in \mathbb{S}_+^{l \times l} \) is a symmetric matrix, it is well known that
\[ \| y_n^i \nabla^2 \hat{f}_n(x_n) y_n^i \|_2 \leq \| \nabla^2 \hat{f}_n(x_n) \|_2 \| y_n^i \|_2^2 \]
for each \( n \in \mathbb{N} \) and \( i \in \{1, \ldots, M_n\} \). Using \( \| y_n^i \|_2 > 0 \) for each \( i = 1, \ldots, M_n \) and \( n \in \mathbb{N} \), we get
\[ \frac{\| y_n^i \nabla^2 \hat{f}_n(x_n) y_n^i \|}{\|y_n^i\|_2} \leq \frac{\| \nabla^2 \hat{f}_n(x_n) \|_2 \| y_n^i \|_2^2}{\|y_n^i\|_2} = \| \nabla^2 \hat{f}_n(x_n) \|_2 \| y_n^i \|_2 \]
Now, from Assumption A 4.9 we know \( \| \nabla^2 f_n(x_n) \|_2 \leq K_H \) for all \( n \in \mathbb{N} \). Hence,

\[
\begin{align*}
\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n c_n \|_2 & \leq \frac{lK_H}{2} \| (Y_n^T W_n Y_n)^{-1} \|_2 \max_{i \in \{1, \ldots, M_n^l\}} \| y_n^i \|_2 \\
& \quad + \left( \frac{lK_H K_y^l}{2} \right) \lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^l\}} \| y_n^i \|_2 \leq 0
\end{align*}
\]

But we know from Assumption A 4.5, that \( \max_{i \in \{1, \ldots, M_n^l\}} \| y_n^i \|_2 \to 0 \) as \( n \to \infty \). Therefore, using (67) in Lemma 4.5, we get

\[
\lim_{n \to \infty} \left( \frac{lK_H}{2} \right) \| (Y_n^T W_n Y_n)^{-1} \|_2 \max_{i \in \{1, \ldots, M_n^l\}} \| y_n^i \|_2 \leq 0
\]

Now, we know from Proposition P 4.4, that \( \{x_n\}_{n \in \mathbb{N}} \subset \bar{D} \) and \( x_n + y_n^i \subset \bar{D} \) for all \( i \in \{M_n^l + 1, \ldots, M_n\} \) and \( n > M \), where \( \bar{D} \subset \bar{E} \) is compact. Therefore, there must exist \( K_y < \infty \) such that \( \max_{i \in \{M_n^l + 1, \ldots, M_n\}} \| y_n^i \|_2 < K_y \) for all \( n \in \mathbb{N} \). Thus, using (68) in Lemma 4.5 we get

\[
\lim_{n \to \infty} \left( \frac{lK_H K_y^l}{2} \right) \lim_{n \to \infty} \| (Y_n^T W_n Y_n)^{-1} \|_2 \max_{i \in \{1, \ldots, M_n^l\}} \| y_n^i \|_2 \leq 0
\]

Therefore \( \| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n c_n \|_2 \to 0 \) as \( n \to \infty \). Hence we have shown that,

\[
\lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 = 0
\]

In particular, if \( x_n \to \bar{x} \in \mathcal{X} \) as \( n \to \infty \), then we get from the continuous differentiability of \( f \) that

\[
\lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(\bar{x}) \right\|_2 \leq \lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 + \lim_{n \to \infty} \left\| \nabla f(x_n) - \nabla f(\bar{x}) \right\|_2 = 0
\]

\[
\blacksquare
\]

**Corollary 4.8.** Consider any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) and let all the assumptions of Theorem 4.7 hold. Suppose for each \( n \in \mathbb{N} \), \( \hat{\nabla}_n f(x_n) \in \mathbb{R}^l \) and \( \hat{\nabla}^2_n f(x_n) \in \mathbb{S}^{l \times l} \) are picked to satisfy (56). Then,

\[
\lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 = 0
\]

**Proof.** The result follows from the fact that if \( \hat{\nabla}_n f(x_n) \) and \( \hat{\nabla}^2_n f(x_n) \) satisfy (56) for each \( n \in \mathbb{N} \), then they also automatically satisfy (58). \( \blacksquare \)
Corollary 4.9. Consider any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) and let all the assumptions of Theorem 4.7 hold. Suppose for each \( n \in \mathbb{N} \), \( \hat{\nabla}_n f(x_n) \in \mathbb{R}^l \) is picked such that
\[
(Y_n^T W_n) Y_n \hat{\nabla}_n f(x_n) = Y_n^T W_n \hat{f}(x_n, Y_n, N_n)
\]
Then we have,
\[
\lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 = 0
\]

Proof. The required result follows from Theorem 4.7 when we set \( \hat{\nabla}_n^2 f(x_n) = 0 \) in (58), for each \( n \in \mathbb{N} \).

In Theorem 4.7, although we used quadratic approximations \( \hat{f}_n \), we only showed the convergence of the approximate gradients \( \nabla \hat{f}_n(x_n) \). Under some stronger conditions we can also show the convergence of the approximate Hessian \( \nabla^2 \hat{f}_n(x_n) \). Next we state the assumptions we need to show that
\[
\lim_{n \to \infty} \left\| \nabla \hat{f}_n(x_n) - \nabla f(x_n) \right\|_2 = 0 \quad \text{and} \quad \lim_{n \to \infty} \left\| \nabla^2 \hat{f}_n(x_n) - \nabla^2 f(x_n) \right\|_2 = 0
\]

As we did earlier, we start by imposing a condition on the convergence of \( \hat{f} \) to \( f \) as the approximation parameter \( N \to \infty \).

A 4.10. For any compact set \( D \subset \mathcal{E} \), we get \( f \in \mathbb{C}^2(D) \), \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \subset \mathbb{W}_1(D) \) and
\[
\lim_{n \to \infty} \left\| \hat{f}(\cdot, N) - f \right\|_{\mathbb{W}_1(D)} = 0
\]

Note that although we assume that \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \subset \mathbb{W}_1(D) \) for any compact \( D \subset \mathcal{E} \), we will not require computation of \( \nabla \hat{f}(x, N) \) for any \( x \) and \( N \).

As in Theorem 4.7, we will require Assumptions A 4.4 through A 4.6 to hold unchanged. Next, let us consider Assumption A 4.7. We might be tempted to impose a condition on the matrices \( (Z^I_n)^T W_n^I Z^I_n \) and \( (Z^O_n)^T W_n^O Z^O_n \) analogous to those imposed on \( (Y_n^I)^T W_n^I Y_n^I \) and \( (Y_n^O)^T W_n^O Y_n^O \) in Assumption A 4.6. However the following result shows that such a condition would be meaningless.

Proposition 4.10. Consider the non-singular regression matrix
\[
Z = \begin{pmatrix}
(z^1)^T \\
\vdots \\
(z^m)^T
\end{pmatrix} \in \mathbb{R}^{m \times (l+1)(l+1)/2}
\]
where for each $i \in \{1, \ldots, m\}$, $(z^i)^T = [(y^i)^T ([y^i]^Q)^T]$ for $y^i \in \mathbb{R}^l$ and $[y^i]^Q$ a vector corresponding to $y^i$ as defined in (46). Also, let $\{w^i : i = 1, \ldots, m\}$ be the set of positive weights associated with the rows $\{(z^i)^T : i = 1, \ldots, m\}$ of $Z$ and let $W = \text{diag}(w^1, \ldots, w^m)$. Without loss of generality, assume that $w^1 \|y^1\|_2^2 \geq w^i \|y^i\|_2^2$ for all $i \in \{1, \ldots, m\}$ and $\delta > 0$ satisfies $\|y^i\|_2 \leq \delta$ for each $i = 1, \ldots, m$. Then,

$$
\|(Z^T W Z)^{-1}\|_2 \|Z^T W Z\|_2 \geq \frac{2}{m \|y^1\|_2^2}
$$

Proof. Since $Z^T W Z$ is a symmetric positive definite matrix, $\|Z^T W Z\|_2 = \lambda_{\max}(Z^T W Z)$ and $\|(Z^T W Z)^{-1}\|_2 = 1/\lambda_{\min}(Z^T W Z)$. Recall that

$$
\lambda_{\max}(Z^T W Z) = \max_{\{z \in \mathbb{R}^{l+(l+1)/2} : \|z\|_2 = 1\}} z^T (Z^T W Z) z
$$

In particular, consider

$$
\tilde{z} = \begin{pmatrix} (y^1)^T \left[ \begin{array}{c} 0, 0, \ldots, 0 \end{array} \right]_{\frac{l(l+1)}{2} \text{ terms}} \end{pmatrix}^T
$$

and note that

$$
Z^T W Z = \sum_{i=1}^{m} w^i (z^i(z^i)^T)
$$

Thus

$$
\lambda_{\max}(Z^T W Z) \geq \tilde{z}^T (Z^T W Z) \tilde{z} = \sum_{i=1}^{m} w^i (\tilde{z}^T z^i)^2 \geq w^1 (\tilde{z}^T z^1)^2 = w^1 \|y^1\|_2^2
$$

Similarly,

$$
\lambda_{\min}(Z^T W Z) = \min_{\{z \in \mathbb{R}^{l+(l+1)/2} : \|z\|_2 = 1\}} z^T (Z^T W Z) z
$$

Consider

$$
\tilde{z} = \begin{pmatrix} 0, 0, \ldots, 0, \frac{[y^1]^Q}{\|y^1\|_2^2} \end{pmatrix}_{l \text{ terms}}
$$

49
Then,
\[
\lambda_{\text{min}}(Z^T W Z) \leq \tilde{z}^T (Z^T W Z) \tilde{z} = \sum_{i=1}^{m} w^i (\tilde{z}^T z^i)^2 = \sum_{i=1}^{m} w^i \left( \frac{[y^i]^T [y^i]^Q}{\| [y^i]^Q \|_2} \right)^2 \leq \sum_{i=1}^{m} w^i \| [y^i]^Q \|_2^2 = \frac{1}{2} \sum_{i=1}^{m} w^i \| y^i \|_2^4 \quad \text{from (47)}
\]
\[
\leq \frac{m}{2} w^1 \| y^1 \|_2^2 \delta^2.
\]

Therefore,
\[
\|(Z^T W Z)^{-1}\|_2 \|Z^T W Z\|_2 = \frac{\lambda_{\text{max}}(Z^T W Z)}{\lambda_{\text{min}}(Z^T W Z)} \geq \frac{2 w^1 \| y^1 \|_2^2}{m w^1 \| y^1 \|_2^2 \delta^2} = \frac{2}{m \delta^2}.
\]

Proposition 4.10 shows that for any sequence \( \{Z^I_n\}_{n \in \mathbb{N}} \) of regression matrices satisfying Assumption A 4.5, and such that the sequence \( \{M^I_n\}_{n \in \mathbb{N}} \) (the number of inner perturbation vectors) remains bounded, the condition number of \((Z^I_n)^T W^I_n Z^I_n\) increases without bound. Thus, in order to state a modified form Assumption A 4.7, we first develop the required notation. For each \( i = 1, \ldots, M_n \) and \( n \in \mathbb{N} \), the scaled perturbation vector \( \tilde{y}^i_n \in \mathbb{R}^l \) and the scaled regression vector \( \tilde{z}^i_n \in \mathbb{R}^p \) be given by
\[
\tilde{y}^i_n := \frac{y^i_n}{\delta_n} \quad \text{and} \quad \tilde{z}^i_n := \left( \frac{y^i_n}{\delta_n} \right)
\]

The corresponding scaled perturbation and regression matrices are defined as follows.
\[
\tilde{Y}_n := \begin{pmatrix} \{y^1_n\}^T \\ \vdots \\ \{y^M_n\}^T \end{pmatrix} = \frac{Y_n}{\delta_n},
\]
\[
\tilde{Y}^Q_n := \begin{pmatrix} \{y^1_n^Q\}^T \\ \vdots \\ \{y^M_n^Q\}^T \end{pmatrix} = \frac{Y^Q_n}{\delta^2_n},
\]
\[
\tilde{Z}_n := \begin{pmatrix} \tilde{Y}_n \\ \tilde{Y}^Q_n \end{pmatrix} = Z_n D_n.
\]

50
where

\[
D_n = \text{diag} \left( \frac{1}{\delta_n}, \ldots, \frac{1}{\delta_n}, \frac{1}{\delta_2^n}, \ldots, \frac{1}{\delta_2^n} \right)
\]

(89)

The corresponding scaled matrices \(\tilde{Y}^I_n, \tilde{Y}^O_n, (\tilde{Y}^I_n)^Q, (\tilde{Y}^O_n)^Q, \tilde{Z}^I_n\) and \(\tilde{Z}^O_n\) are defined analogously.

Now, we are ready to state our modifications to Assumption A 4.7.

A 4.11. The set of design points \(\{x_n + y_n^i : i = 1, \ldots, M_n\text{ and } n \in \mathbb{N}\}\) and the sequence \(\{W_n\}_{n \in \mathbb{N}}\) of weight matrices, satisfy the following.

1. There exists \(K^I_\lambda < \infty\) such that for each \(n \in \mathbb{N}\)

\[
\left\| (\tilde{Z}^I_n)^T W_n^I \tilde{Z}^I_n \right\|_2 \left\| (\tilde{Z}^I_n)^T W_n^I \tilde{Z}^I_n \right\|^{-1}_2 < K^I_\lambda
\]

(90)

2.

\[
\lim_{n \to \infty} \left\| (\tilde{Z}^O_n)^T W_n^O \tilde{Z}^O_n \right\|_2 = 0
\]

(91)

We make a few comments regarding Assumption A (4.11).

- Let \((\tilde{z}^i_n)^T\) denote row \(i\) of \(\tilde{Z}_n\). Then

\[
\tilde{Z}_n = \begin{pmatrix}
(\tilde{z}^1_n)^T \\
\vdots \\
(\tilde{z}^{M_n}_n)^T
\end{pmatrix} = Z_n D_n = \begin{pmatrix}
\frac{1}{\delta_n} (y_n^1)^T \\
\vdots \\
\frac{1}{\delta_n} (y_n^{M_n})^T
\end{pmatrix}
\]

\[
\begin{pmatrix}
\frac{1}{\delta_n} ([y_n^1]^Q)^T \\
\vdots \\
\frac{1}{\delta_n} ([y_n^{M_n}]^Q)^T
\end{pmatrix}
\]

Thus,

\[
\tilde{z}^i_n = \begin{pmatrix}
y_n^i \\
\frac{y_n^i}{\delta_n} \\
[y_n^i]^Q \frac{y_n^i}{\delta_n}
\end{pmatrix}
\]

for each \( n \in \mathbb{N} \text{ and } i \in \{1, \ldots, M_n\}\)

Hence,

\[
\left\| \tilde{z}^i_n \right\|^2_2 = \left\| y_n^i \right\|^2_2 + \left\| [y_n^i]^Q \right\|^2_2 \frac{1}{\delta_n^4} = \left( \left\| y_n^i \right\|_2 \right)^2 \frac{1}{\delta_n^4} + \frac{1}{2} \left( \left\| y_n^i \right\|_2 \right)^4
\]

(92)

- Note that since \(\delta_n > 0\) for each \(n \in \mathbb{N}\), \(D_n\) is non-singular for each \(n \in \mathbb{N}\). Hence

\[
(Z_n^T W_n Z_n)^{-1} Z_n^T = D_n(D_n Z_n^T W_n Z_n D_n)^{-1} (Z_n D_n)^T = D_n (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \tilde{Z}_n^T
\]

(93)
From Assumption A 4.11 we get the following result analogous to Lemma 4.5.

**Lemma 4.11.** Let Assumption A 4.11 hold. Then, \(\hat{Z}_n^TW_n\hat{Z}_n\) and \(Z_n^O W_n Z_n\) are positive definite for each \(n \in \mathbb{N}\) and the following hold true.

1. 
\[
\left\| (\hat{Z}_n^TW_n\hat{Z}_n)^{-1} \right\|_2 \left\| (\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I \right\|_2 < K_\lambda^I \quad \text{for all } n \in \mathbb{N} \quad (94)
\]

2. 
\[
\lim_{n \to \infty} \left\| (\hat{Z}_n^TW_n\hat{Z}_n)^{-1} \right\|_2 \left\| (\hat{Z}_n^O)^T W_n^O \hat{Z}_n^O \right\|_2 = 0 \quad (95)
\]

**Proof.** Since \((\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I\) is a symmetric and positive semidefinite matrix, we get from (90) that
\[
\lambda_{\min}( (\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I ) = \frac{1}{\left\| (\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I \right\|_2^{-1}} > 0
\]
Further, it is easily seen that 
\[
\hat{Z}_n^TW_n\hat{Z}_n = (\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I + (\hat{Z}_n^O)^T W_n^O \hat{Z}_n^O
\]
Now, since \((\hat{Z}_n^O)^T W_n^O Z_n^O\) is positive semidefinite, from the interlocking eigenvalues theorem, we know that 
\[
\lambda_{\min}(\hat{Z}_n^TW_n\hat{Z}_n) \geq \lambda_{\min}( (\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I ) > 0
\]
Therefore, \(\hat{Z}_n^TW_n\hat{Z}_n\) is positive definite. Further, from the definition of \(D_n\) in (89), we know that \(D_n\) is positive definite for each \(n \in \mathbb{N}\). Since the product of positive definite matrices remains positive definite, we get that 
\[
Z_n^TW_nZ_n = D_n^{-1}(\hat{Z}_n^TW_n\hat{Z}_n)D_n^{-1}
\]
is positive definite for each \(n \in \mathbb{N}\). Now,
\[
\left\| (\hat{Z}_n^TW_n\hat{Z}_n)^{-1} \right\|_2 = \frac{1}{\lambda_{\min}(\hat{Z}_n^TW_n\hat{Z}_n)} \leq \frac{1}{\lambda_{\min}((\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I)} = \left\| ((\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I)^{-1} \right\|_2 \quad (96)
\]
Therefore, we see that for any \(n \in \mathbb{N}\), using (96),
\[
\left\| (\hat{Z}_n^TW_n\hat{Z}_n)^{-1} \right\|_2 \left\| (\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I \right\|_2 \leq \left\| ((\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I)^{-1} \right\|_2 \left\| (\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I \right\|_2 \leq K_\lambda^I
\]
Thus, we have shown that (94) holds. Similarly,
\[
\left\| (\hat{Z}_n^TW_n\hat{Z}_n)^{-1} \right\|_2 \left\| (\hat{Z}_n^O)^T W_n^O \hat{Z}_n^O \right\|_2 \leq \left\| (\hat{Z}_n^TW_n\hat{Z}_n)^{-1} \right\|_2 \left\| (\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I \right\|_2 \left( \frac{\left\| (\hat{Z}_n^O)^T W_n^O \hat{Z}_n^O \right\|_2}{\left\| (\hat{Z}_n^I)^T W_n^I \hat{Z}_n^I \right\|_2} \right)
\]

52
Using (94) on the right hand side of the above equation, we get

\[ \| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \|_2 \| (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \|_2 \leq K^I_\lambda \left( \| (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \|_2 \right) \]

Finally, we use (64) in Assumption A 4.11 to get that

\[ \lim_{n \to \infty} \| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \|_2 \| (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \|_2 \leq K^I_\lambda \lim_{n \to \infty} \| (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \|_2 = 0 \]

Finally, we use the following stronger version of Assumption A 4.8.

A 4.12. There exists a sequence \( \{ N_n^\# \}_{n \in \mathbb{N}} \) some \( K^\# \in \mathbb{N} \) such that

1. \( N_n^i = N_n^\# \) for all \( i = 1, \ldots, M_n \) and \( n > K^\# \).
2. \( N_n^\# \to \infty \) as \( n \to \infty \).

From Assumption A 4.10 we know that \( \nabla^2 f(x) \) exists for all \( x \in \mathcal{E} \). Therefore, analogous to the definition of \( [\hat{\nabla}^2 f_n(x)]_v \) in (53), we define vector \( [\nabla^2 f(x)]_v \in \mathbb{R}^{l(l+1)/2} \) for any \( x \in \mathcal{E} \) containing the components of \( \nabla^2 f(x) \) as follows:

\[ [\nabla^2 f(x)]_v := \left\{ [\nabla^2 f(x)]_{jj} \right\}_{j=1, \ldots, l}, \left\{ [\nabla^2 f(x)]_{jk} \right\}_{j,k=1, \ldots, l, j < k} \] (97)

where \( [\nabla^2 f(x)]_{jk} \) denotes the element of \( \nabla^2 f(x) \) in row \( j \) and column \( k \). With these assumptions, we are ready to state and prove our convergence result for \( \hat{\nabla} f_n(x_n) \) and \( \hat{\nabla}^2 f_n(x_n) \).

Theorem 4.12. Consider a sequence \( \{ x_n \}_{n \in \mathbb{N}} \subset \mathcal{D} \subset \mathcal{E} \) where \( \mathcal{D} \) is compact and let for each \( n \in \mathbb{N} \), \( \hat{\nabla}_n f(x_n) \in \mathbb{R}^l \) and \( \hat{\nabla}^2_n f(x_n) \in \mathbb{S}^{l \times l} \) be picked so as to satisfy (56). Suppose that Assumptions A 4.4 through A 4.6 and Assumptions A 4.10 through A 4.12 hold. Then, we have,

\[ \lim_{n \to \infty} \| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \|_2 = 0 \quad \text{and} \quad \lim_{n \to \infty} \| \hat{\nabla}^2_n f(x_n) - \nabla^2 f(x_n) \|_2 = 0 \]

In particular, if \( \{ x_n \}_{n \in \mathbb{N}} \subset \mathcal{D} \) is such that \( x_n \to \bar{x} \in \mathcal{D} \) as \( n \to \infty \), then

\[ \lim_{n \to \infty} \hat{\nabla}_n f(x_n) = \nabla f(\bar{x}) \quad \text{and} \quad \lim_{n \to \infty} \hat{\nabla}^2_n f(x_n) = \nabla^2 f(\bar{x}) \]

53
Proof of Theorem 4.12:

From Lemma 4.11, we know that \( Z_n^T W_n Z_n \) is positive definite and hence non-singular for each \( n \in \mathbb{N} \). Therefore, we can rewrite (56) as

\[
\begin{pmatrix}
\hat{\nabla}_n f(x_n) \\
[\nabla^2 f(x_n)]_v
\end{pmatrix}
-
\begin{pmatrix}
\nabla f(x_n) \\
[\nabla^2 f(x_n)]_v
\end{pmatrix}
= (Z_n^T W_n Z_n)^{-1} Z_n^T W_n \hat{f}(x_n, Y_n, N_n) - (Z_n^T W_n Z_n)^{-1} Z_n^T W_n Z_n \begin{pmatrix}
\nabla f(x_n) \\
[\nabla^2 f(x_n)]_v
\end{pmatrix}
\]

Noting from Assumption A 4.10 that \( \hat{f}(\cdot, N) \in \mathbb{W}_1(D) \) for each \( N \in \mathbb{N} \) and since \( \{x_n\}_{n \in \mathbb{N}} \subset D \), we define the vector \( b_n \in \mathbb{R}^{M_n} \) as,

\[
b_n := \begin{pmatrix}
y_1^T(\nabla \hat{f}(x_n, N_n^1) - \nabla f(x_n)) \\
\vdots \\
y_{M_n}^T(\nabla \hat{f}(x_n, N_n^{M_n}) - \nabla f(x_n))
\end{pmatrix}
\]

Adding and subtracting \( (Z_n^T W_n Z_n)^{-1} Z_n^T W_n (b_n + f(x_n, Y_n)) \) in (98), we get

\[
\begin{pmatrix}
\hat{\nabla}_n f(x_n) \\
[\nabla^2 f(x_n)]_v
\end{pmatrix}
-
\begin{pmatrix}
\nabla f(x_n) \\
[\nabla^2 f(x_n)]_v
\end{pmatrix}
= (Z_n^T W_n Z_n)^{-1} Z_n^T W_n \left[ \hat{f}(x_n, Y_n, N_n) - f(x_n, Y_n) - b_n \right] +
\]

\[
(Z_n^T W_n Z_n)^{-1} Z_n^T W_n b_n + (Z_n^T W_n Z_n)^{-1} Z_n^T W_n \left[ f(x_n, Y_n) - Z_n \begin{pmatrix}
\nabla f(x_n) \\
[\nabla^2 f(x_n)]_v
\end{pmatrix}
\right]
\]

Using the fact that \( (Z_n^T W_n Z_n)^{-1} Z_n^T W_n = D_n(\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n \), we get

\[
\begin{pmatrix}
\hat{\nabla}_n f(x_n) \\
[\nabla^2 f(x_n)]_v
\end{pmatrix}
-
\begin{pmatrix}
\nabla f(x_n) \\
[\nabla^2 f(x_n)]_v
\end{pmatrix}
= D_n(\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n \left[ \hat{f}(x_n, Y_n, N_n) - f(x_n, Y_n) - b_n \right] +
\]

\[
D_n(\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n b_n + D_n(\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n \left[ f(x_n, Y_n) - Z_n \begin{pmatrix}
\nabla f(x_n) \\
[\nabla^2 f(x_n)]_v
\end{pmatrix}
\right]
\]

From Assumption A 4.4 we know that \( \|y_i^k\|_2 > 0 \) for all \( i \in \{1, \ldots, M_n\} \) and \( n \in \mathbb{N} \). Using (47), it is easily seen that this means \( \|z_i^k\|_2 > 0 \) for all \( i \in \{1, \ldots, M_n\} \) and \( n \in \mathbb{N} \). Therefore, we set for each \( n \in \mathbb{N} \),

\[
a_n := \left( \hat{f}(x_n, Y_n, N_n) - f(x_n, Y_n) - b_n \right)
= \begin{pmatrix}
\|z_1^k\|_2 \left( \hat{f}(x_n+y_1^k, N_n^1) - f(x_n+N_n^1) \right) - y_1^k (\nabla \hat{f}(x_n, Y_n) - \nabla f(x_n)) \\
\vdots \\
\|z_{M_n}^k\|_2 \left( \hat{f}(x_n+y_{M_n}^k, N_n^{M_n}) - f(x_n+N_n^{M_n}) \right) - y_{M_n}^k (\nabla \hat{f}(x_n, Y_n) - \nabla f(x_n))
\end{pmatrix}
\]
Let us simplify the third term on the right side of (100).

\[
f(x_n, Y_n) - Z_n \left( \frac{\nabla f(x_n)}{[\nabla^2 f(x_n)]_v} \right) = f(x_n, Y_n) - Y_n \nabla f(x_n) - Y_n^Q \nabla^2 f(x_n)
\]

Again noting that for each \( n \in \mathbb{N} \) and \( \|z_i^n\|_2 > 0 \) for each \( i \in \{1, \ldots, M_n\} \) and \( n \in \mathbb{N} \), we set

\[
c_n := f(x, Y_n) - Z_n \left( \frac{\nabla f(x_n)}{[\nabla^2 f(x_n)]_v} \right) = \begin{pmatrix}
\|z_1^n\|_2 f(x_n + y_1^n) - f(x_n) - y_1^n T \nabla f(x_n) - \frac{1}{2} (y_1^n T \nabla^2 f(x_n) y_1^n)
\|z_2^n\|_2 f(x_n + y_2^n) - f(x_n) - y_2^n T \nabla f(x_n) - \frac{1}{2} (y_2^n T \nabla^2 f(x_n) y_2^n)
\|z_{M_n}^n\|_2 f(x_n + y_{M_n}^n) - f(x_n) - y_{M_n}^n T \nabla f(x_n) - \frac{1}{2} (y_{M_n}^n T \nabla^2 f(x_n) y_{M_n}^n)
\end{pmatrix}
\]

Therefore, we finally have for each \( n \in \mathbb{N} \),

\[
\begin{pmatrix}
\hat{\nabla}_n f(x_n)
[\nabla^2 f(x_n)]_v
\end{pmatrix} - \begin{pmatrix}
\nabla f(x_n)
[\nabla^2 f(x_n)]_v
\end{pmatrix} = D_n (\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n a_n + D_n (\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n b_n + D_n (\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n c_n
\]

\[
\Rightarrow \left\| \begin{pmatrix}
\hat{\nabla}_n f(x_n)
[\nabla^2 f(x_n)]_v
\end{pmatrix} - \begin{pmatrix}
\nabla f(x_n)
[\nabla^2 f(x_n)]_v
\end{pmatrix} \right\|_2 \leq \left\| D_n (\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n a_n \right\|_2 + \left\| D_n (\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n b_n \right\|_2 + \left\| D_n (\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n c_n \right\|_2
\]

(101)

Next, we show that all three terms on the right side of (101) converge to zero as \( n \to \infty \). But first, we note that from Assumption A 4.5, \( \delta_n \to 0 \) as \( n \to \infty \). Therefore, there exists \( K^* \in \mathbb{N} \), such that for all \( n > K^* \), \( \delta_n < 1 \). This means that for all \( n > K^* \), \((1/\delta_n)^2 > (1/\delta_n)\). Therefore, we get that for all \( n > K^* \), \( \|D_n\|_2 = (1/\delta_n)^2 \). For the remainder of our proof we will assume that we consider only \( n > K^* \).

55
Let us consider the first term right side of (101) for \( n > K^* \). First, we note that
\[
\| D_n (\tilde{Z}^T_n W_n \tilde{Z}_n)^{-1} \tilde{Z}^T_n W_n a_n \|_2 \leq \| D_n \|_2 \left\| (\tilde{Z}^T_n W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| \tilde{Z}^T_n W_n a_n \right\|_2 = \left( \frac{1}{\delta^2_n} \right) \left\| (\tilde{Z}^T_n W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| \tilde{Z}^T_n W_n a_n \right\|_2
\]

Recall that the matrix \( \tilde{Z}_n \) has \( p = l + (l + 1)/2 \) columns. Using this fact in (71) of Lemma 4.6, we get
\[
\max_{i \in \{1,...,M^*_n\}} \left\| \hat{f}(x_n + y^i_n, N^i_n) - \hat{f}(x_n, N^i_n) \right\| \leq \frac{\| (\tilde{Z}^T_n W_n \tilde{Z}_n)^{-1} \|_2}{\delta^2_n \| \tilde{z}^i_n \|_2} \times \left\{ \| (\tilde{Z}^T_n I^I_n W_n \tilde{Z}_n)^{-1} \|_2 \times \left\| \hat{f}(x_n + y^i_n, N^i_n) - \hat{f}(x_n, N^i_n) \right\| + \left\| (\tilde{Z}^O_n I^O_n W_n \tilde{Z}_n)^{-1} \|_2 \times \left\| \hat{f}(x_n + y^i_n, N^i_n) - \hat{f}(x_n, N^i_n) \right\| \right\}
\]

From the above, we get
\[
\| D_n (\tilde{Z}^T_n W_n \tilde{Z}_n)^{-1} \tilde{Z}^T_n W_n a_n \|_2 \leq p \left\| (\tilde{Z}^T_n W_n \tilde{Z}_n)^{-1} \right\|_2 \times \left\{ \left\| (\tilde{Z}^T_n I^I_n W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| \hat{f}(x_n + y^i_n, N^i_n) - \hat{f}(x_n, N^i_n) \right\| + \left\| (\tilde{Z}^O_n I^O_n W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| \hat{f}(x_n + y^i_n, N^i_n) - \hat{f}(x_n, N^i_n) \right\| \right\}
\]

First, from Assumption A 4.11 and Lemma 4.11 we get that
\[
\left\| (\tilde{Z}^T_n W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}^T_n I^I_n W_n \tilde{Z}_n)^{-1} \right\|_2 \leq K^I_n \quad \text{for all } n \in \mathbb{N}
\]
and
\[
\lim_{n \to \infty} \left\| (\tilde{Z}^T_n W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}^O_n I^O_n W_n \tilde{Z}_n)^{-1} \right\|_2 = 0
\]
Therefore, clearly there exists a constant \( K_\lambda < \infty \) such that for all \( n \in \mathbb{N} \),
\[
\left\| (\tilde{Z}^T_n W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}^T_n I^I_n W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}^O_n I^O_n W_n \tilde{Z}_n)^{-1} \right\|_2 < K_\lambda
\]

Next, we note that from (47), for any \( n \in \mathbb{N} \) and \( i \in \{1,...,M^*_n\} \),
\[
\delta^2_n \| \tilde{z}^i_n \|_2 = \sqrt{\delta^2_n \| y^i_n \|_2^2 + \left( \frac{1}{2} \right) \| y^i_n \|_2^4} \geq \left( \frac{\sqrt{1}}{2} \right) \| y^i_n \|_2^2
\]
Therefore, for each \( i = 1, \ldots, M_n \),
\[
\frac{\left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - \left( f(x_n + y_n^i) - f(x_n) \right) \right|}{\|y_n^i\|_2^2} \leq \beta_n \|z_n\|_2
\]
\[
\frac{\sqrt{2} \left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - \left( f(x_n + y_n^i) - f(x_n) \right) \right|}{\|y_n^i\|_2^2} \leq \beta_n \|z_n\|_2
\]

First, from Proposition P 4.4, we know that there exists \( \hat{D} \subset E \) such that \( \{x_n\}_{n \in \mathbb{N}} \subset \hat{D} \) and there exists \( M \in \mathbb{N} \) such that for all \( n > M, x_n + y_n^i \in \hat{D} \) and \( i = 1, \ldots, M_n \). Also, from Assumption A 4.10, we know that \( \|\hat{f}(\cdot, N) - f\|_{W_1(\hat{D})} \to 0 \) as \( N \to \infty \). Thus, using Lemma 2.5 we get that for any \( \varepsilon > 0 \) there exists \( N_\varepsilon \in \mathbb{N} \) such that for all \( N > N_\varepsilon \),
\[
\sup_{x, x + y \in \hat{D}} \frac{|\hat{f}(x + y, N) - f(x + y) - (\hat{f}(x, N) - f(x)) - y^T(\nabla \hat{f}(x, N) - \nabla f(x))|}{\|y\|_2^2} < \varepsilon
\]

Now from Assumption A 4.8, we get that, for any \( N_\varepsilon \in \mathbb{N} \), there exists \( K_s \in \mathbb{N} \) such that for all \( n > K_s, N_n^i > N_\varepsilon \) for each \( i = 1, \ldots, M_n \). Therefore, combining these two observations, we get that given any \( \varepsilon > 0 \), there exists some \( K_s > M \) such that for all \( n > \max\{K_s, K^*\} \),
\[
\max_{i \in \{1, \ldots, M_n\}} \frac{|\hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) - (f(x_n + y_n^i) - f(x_n)) - y_n^i T(\nabla \hat{f}(x_n, N_n^i) - \nabla f(x_n))|}{\|y_n^i\|_2^2} \leq \varepsilon
\]
\[
\sup_{x, x + y \in \hat{D}} \frac{|\hat{f}(x + y, N_n^i) - f(x + y) - (\hat{f}(x, N_n^i) - f(x)) - y^T(\nabla \hat{f}(x, N_n^i) - \nabla f(x))|}{\|y\|_2^2} < \varepsilon
\]

Therefore, we get that
\[
\lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n\}} \frac{|\hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) - (f(x_n + y_n^i) - f(x_n)) - y_n^i T(\nabla \hat{f}(x_n, N_n^i) - \nabla f(x_n))|}{\|y_n^i\|_2^2} = 0
\]

And hence we finally get
\[
\lim_{n \to \infty} \left\| D_n(\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \tilde{Z}_n^T W_n a_n \right\|_2 \leq \sqrt{2} p \ K_\lambda \times
\]
\[
\lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n\}} \frac{|\hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) - (f(x_n + y_n^i) - f(x_n)) - y_n^i T(\nabla \hat{f}(x_n, N_n^i) - \nabla f(x_n))|}{\|y_n^i\|_2^2} = 0
\]

Next, let us consider the second term on the right side of (101). Letting \( 0_{l(l+1)/2} \) denote a column vector with \( l(l+1)/2 \) elements each of which is equal to zero, it is easily seen from Assumption
A.4.12 the for all \( n > K^\# \),

\[
\begin{align*}
    b_n &= \left( \begin{array}{c}
y_n^T(\nabla \hat{f}(x_n, N_n^1) - \nabla f(x_n)) \\
    \vdots \\
    y_n^{M_n}(\nabla \hat{f}(x_n, N_n^{M_n}) - \nabla f(x_n))
\end{array} \right) = \left( \begin{array}{c}
y_n^T(\nabla \hat{f}(x_n, N_n^{\#}) - \nabla f(x_n)) \\
    \vdots \\
    y_n^{M_n}(\nabla \hat{f}(x_n, N_n^{\#}) - \nabla f(x_n))
\end{array} \right) \\
    &= Y_n^T(\hat{f}(x_n, N_n^{\#}) - \nabla f(x_n)) \\
    &= Z_n^T\left(\hat{f}(x_n, N_n^{\#}) - \nabla f(x_n)\right) \\
    &= Z_n^TD_n^{-1}\left(\hat{f}(x_n, N_n^{\#}) - \nabla f(x_n)\right)
\end{align*}
\]

Clearly,

\[
\|D_n(Z_n^TW_n\hat{Z}_n)^{-1}Z_n^TW_nb_n\|_2 \leq \|D_n\|_2 \|(Z_n^TW_n\hat{Z}_n)^{-1}\|_2 \|Z_n^TW_nb_n\|_2 = \left(\frac{1}{\delta_n}^2\right)\|(Z_n^TW_n\hat{Z}_n)^{-1}\|_2 \|Z_n^TW_nb_n\|_2
\]

Therefore, using the definition of \( b_n \), we get for each \( n > K^\# \)

\[
\|D_n(Z_n^TW_n\hat{Z}_n)^{-1}Z_n^TW_nb_n\|_2 = \|D_n(Z_n^TW_n\hat{Z}_n)^{-1}Z_n^TW_n\hat{Z}_nD_n^{-1}\left(\nabla \hat{f}(x_n, N_n^{\#}) - \nabla f(x_n)\right)\|_2 = \|\nabla \hat{f}(x_n, N_n^{\#}) - \nabla f(x_n)\|_2
\]

Since \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D} \subset \mathcal{E} \) where \( \mathcal{D} \) is compact, we get from Assumption A.4.10 \( \|\hat{f}(\cdot, N) - f\|_{W_1(\mathcal{D})} \to 0 \) as \( N \to \infty \). Further, we know from Assumption A.4.12 that \( N_n^{\#} \to \infty \) as \( n \to \infty \). Therefore, we get from the definition of \( \|\cdot\|_{W_1} \) that

\[
\lim_{n \to \infty} \|\nabla \hat{f}(x_n, N_n^{\#}) - \nabla f(x_n)\|_2 = 0
\]

Therefore,

\[
\lim_{n \to \infty} \|D_n(Z_n^TW_n\hat{Z}_n)^{-1}Z_n^TW_nb_n\|_2 = 0
\]

Finally, we consider the third term on the right side of (101) for \( n > K^* \).

\[
\|D_n(Z_n^TW_n\hat{Z}_n)^{-1}Z_n^TW_n\hat{Z}_n\|_2 \leq \|D_n\|_2 \|(Z_n^TW_n\hat{Z}_n)^{-1}\|_2 \|Z_n^TW_n\hat{Z}_n\|_2 = \left(\frac{1}{\delta_n}^2\right)\|(Z_n^TW_n\hat{Z}_n)^{-1}\|_2 \|Z_n^TW_n\hat{Z}_n\|_2
\]
Using Lemma 4.6 and the definition of $c_n$, we get

\[
\begin{align*}
\|D_n(\tilde{Z}_n W_n \tilde{Z}_n)^{-1} \tilde{Z}_n W_n c_n\|_2 & \leq p \|\tilde{Z}_n W_n \tilde{Z}_n\|_2^{-1} \times \\
\left\{ \left\| (\tilde{Z}_n^I W_n^I \tilde{Z}_n^I) \right\|_{\max_{2 \in \{1, \ldots, M_n^I\}}} \right\} & + \\
\left\{ \left\| (\tilde{Z}_n^O W_n^O \tilde{Z}_n^O) \right\|_{\max_{2 \in \{M_n^O+1, \ldots, 1\}}} \right\} & + \\
\end{align*}
\]

Again using (103), we get that

\[
\begin{align*}
\|D_n(\tilde{Z}_n W_n \tilde{Z}_n)^{-1} \tilde{Z}_n W_n c_n\|_2 & \leq \sqrt{2} p \|\tilde{Z}_n W_n \tilde{Z}_n\|_2^{-1} \times \\
\left\{ \left\| (\tilde{Z}_n^I W_n^I \tilde{Z}_n^I) \right\|_{\max_{2 \in \{1, \ldots, M_n^I\}}} \right\} & + \\
\left\{ \left\| (\tilde{Z}_n^O W_n^O \tilde{Z}_n^O) \right\|_{\max_{2 \in \{M_n^O+1, \ldots, 1\}}} \right\} & + \\
\end{align*}
\]

(104)

Let us consider the two terms on the right side of (104). First, we show that

\[
\lim_{n \to \infty} \sqrt{2} p \|\tilde{Z}_n W_n \tilde{Z}_n\|_2^{-1} \left\| (\tilde{Z}_n^I W_n^I \tilde{Z}_n^I) \right\|_2 = 0
\]

From Assumption A 4.11 and Lemma 4.11, we know that for all $n \in \mathbb{N}$,

\[
\|\tilde{Z}_n W_n \tilde{Z}_n\|_2^{-1} \left\| (\tilde{Z}_n^I W_n^I \tilde{Z}_n^I) \right\|_2 \leq K_n^I
\]

From Proposition P 4.4 we know that $\{x_n\}_{n \in \mathbb{N}} \subset \tilde{D} \subset \mathcal{E}$ and for all $n > M$, $x_n + y_n^i \in \mathcal{E}$ for $n > M$ and $i = 1, \ldots, M_n$ where $\tilde{D}$ is compact. Also, from Assumption A 4.10, we know that $f \in C_2(\tilde{D})$. Since $\tilde{D} \subset \mathcal{E}$ is compact, $\nabla^2 f$ is uniformly continuous on $\tilde{D}$. That is, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

\[
\sup_{x+y, x \in \tilde{D}, \ |y|_2 \leq \delta} \left| \frac{f(x+y) - f(x) - y^T \nabla f(x) - \frac{1}{2}y^T \nabla^2 f(x)y}{\|y\|_2^2} \right| < \varepsilon
\]

Further, using Assumption A 4.5 we get

\[
\lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^I\}} \|x_n + y_n^i - x_n\|_2 = \lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^I\}} \|y_n^i\|_2 \leq \delta_n = 0
\]

(106)
It follows that for any \( \delta > 0 \), there exists an \( \tilde{K} \in \mathbb{N} \) with such that for all \( n > \tilde{K}, \delta_n < \delta \). Thus, \( 0 < \|y^i_n\|_2 \leq \delta \) for all \( i = 1, \ldots, M^I_n \) and \( n > \tilde{n} \). Thus, we get from (105) that for any \( \varepsilon > 0 \), there exists \( \tilde{K} \in \mathbb{N} \) such that

\[
\left\| f(x_n + y^i_n) - f(x_n) - y^i_n^T \nabla f(x_n) - \frac{1}{2} \left( y^i_n^T \nabla^2 f(x_n)y^i_n \right) \right\|_2 < \varepsilon
\]

for \( i \in \{1, \ldots, M^I_n\} \) and \( n > \max\{\tilde{K}, K^*, M\} \). Consequently,

\[
\lim_{n \to \infty} \max_{i \in \{1, \ldots, M^I_n\}} \left| \frac{f(x_n + y^i_n) - f(x_n) - y^i_n^T \nabla f(x_n) - \frac{1}{2} \left( y^i_n^T \nabla^2 f(x_n)y^i_n \right)}{\|y^i_n\|_2^2} \right| = 0
\]

Therefore, we see that

\[
\lim_{n \to \infty} \sqrt{2} p \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^I T W_n^I \tilde{Z}_n^I)^{\max_{i \in \{1, \ldots, M^I_n\}}} \right\|_2 \left| \frac{f(x_n + y^i_n) - f(x_n) - y^i_n^T \nabla f(x_n) - \frac{1}{2} \left( y^i_n^T \nabla^2 f(x_n)y^i_n \right)}{\|y^i_n\|_2^2} \right| = 0
\]

Finally, we show that

\[
\lim_{n \to \infty} \sqrt{2} p \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)\right\|_2 \times \left( (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \right)_{\max_{i \in \{M^I_n+1, \ldots, M_n\}}} \left| \frac{f(x_n + y^i_n) - f(x_n) - y^i_n^T \nabla f(x_n) - \frac{1}{2} \left( y^i_n^T \nabla^2 f(x_n)y^i_n \right)}{\|y^i_n\|_2^2} \right| = 0
\]

As we noted earlier, we get from Assumption A 4.10, \( f \) is twice continuously differentiable on \( \tilde{D} \).

Further, since \( \tilde{D} \) is compact, we get from Lemma 2.4 that there exists \( K_{2f} < \infty \) such that

\[
\sup_{x, x+y \in \tilde{D}, y \neq 0} \left| \frac{f(x+y) - f(x) - y^T \nabla f(x) - \left( \frac{1}{2} \right) y^T \nabla^2 f(x) y}{\|y\|_2^2} \right| < K_{2f}
\]

Since \( \{x_n + y^i_n : i = 1, \ldots, M_n\} \subset \tilde{D} \) for \( n > M \) and \( \{x_n\}_{n \in \mathbb{N}} \subset \tilde{D} \), it follows that

\[
\max_{i \in \{M^I_n+1, \ldots, M_n\}} \left| \frac{f(x_n + y^i_n) - f(x_n) - y^i_n^T \nabla f(x_n) - \frac{1}{2} \left( y^i_n^T \nabla^2 f(x_n)y^i_n \right)}{\|y^i_n\|_2^2} \right| \leq K_{2f}
\]

Also, from Assumption A 4.11 and Lemma 4.11, we get that

\[
\lim_{n \to \infty} \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \right\|_2 = 0
\]
Therefore, we finally get
\[
\lim_{n \to \infty} \sqrt{2} p \left\| (\tilde{Z}_n^TW_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^OW_n \tilde{Z}_n^O) \right\|_2 \times \max_{i \in \{M_1+1, \ldots, M_n\}} \left| \frac{f(x_n + y_n^i) - f(x_n) - y_n^T \nabla f(x_n) - \frac{1}{2} y_n^T \nabla^2 f(x_n) y_n^i}{\|y_n^i\|_2^2} \right| \leq \sqrt{2} p \ K_2 f \ \lim_{n \to \infty} \left\| (\tilde{Z}_n^TW_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^OW_n \tilde{Z}_n^O) \right\|_2 = 0
\]

Thus, we have shown that \( \left\| D_n(\tilde{Z}_n^TW_n \tilde{Z}_n)^{-1} \tilde{Z}_n^TW_n c_n \right\|_2 \to 0 \) as \( n \to \infty \). Hence it holds that
\[
\lim_{n \to \infty} \left\| \left( \nabla f(x_n) \right) - \left( \nabla f(x_0) \right) \right\|_2 = 0
\]

This in turn gives us
\[
\lim_{n \to \infty} \left\| \tilde{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 = 0 \quad \text{and} \quad \lim_{n \to \infty} \left\| \tilde{\nabla}_n^2 f(x_n) - \nabla^2 f(x_n) \right\|_2 = 0.
\]

In particular, if \( x_n \to \tilde{x} \in D \) as \( n \to \infty \), then since \( f \in C_2(D) \),
\[
\lim_{n \to \infty} \left\| \nabla f(x_n) - \nabla f(\tilde{x}) \right\|_2 \leq \lim_{n \to \infty} \left\| \nabla f(x_n) - \nabla f(x_0) \right\|_2 + \lim_{n \to \infty} \left\| \nabla f(x_0) - \nabla f(\tilde{x}) \right\|_2 = 0
\]
\[
\lim_{n \to \infty} \left\| \tilde{\nabla}_n^2 f(x_n) - \tilde{\nabla}_n^2 f(\tilde{x}) \right\|_2 \leq \lim_{n \to \infty} \left\| \tilde{\nabla}_n^2 f(x_n) - \tilde{\nabla}_n^2 f(x_0) \right\|_2 + \lim_{n \to \infty} \left\| \tilde{\nabla}_n^2 f(x_0) - \tilde{\nabla}_n^2 f(\tilde{x}) \right\|_2 = 0.
\]

\[
\mathbf{5} \quad \text{Iterative Search Algorithms}
\]

In Section 4, we discussed methods to obtain sequences \( \{f_n\}_{n \in \mathbb{N}}, \{\tilde{\nabla}_n f\}_{n \in \mathbb{N}} \) and \( \{\tilde{\nabla}_n^2 f\}_{n \in \mathbb{N}} \) that approximate respectively a function \( f \), its gradient \( \nabla f \) and Hessian \( \nabla^2 f \). In this section, we consider the use of such approximations in the design of iterative search algorithms for the optimization problem \( (P) \).

At \( 5.1 \). The objective function \( f \) is continuously differentiable on a neighborhood of \( \mathcal{X} \).

We consider the case of problem \( (P) \) when neither \( f(x) \) nor \( \nabla f(x) \) can be evaluated exactly for a given \( x \in \mathcal{X} \). Instead, we assume that \( f \) and \( \nabla f \) can be approximated by the sequences \( \{f_n\}_{n \in \mathbb{N}} \) and \( \{\tilde{\nabla}_n f\}_{n \in \mathbb{N}} \) respectively. In particular, unless stated otherwise, we will assume the following regarding \( \{f_n\}_{n \in \mathbb{N}} \) and \( \{\tilde{\nabla}_n f\}_{n \in \mathbb{N}} \).
A 5.2. For any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that \( x_n \to x \) as \( n \to \infty \), it holds that

\[
\lim_{n \to \infty} f_n(x_n) = f(x) \quad \text{and} \quad \lim_{n \to \infty} \nabla_n f(x_n) = \nabla f(x)
\] (107)

Remarks:

- Recall that the sequences \( \{f_n\}_{n \in \mathbb{N}} \) and \( \{\nabla \hat{f}_n\}_{n \in \mathbb{N}} \) obtained in Section 4 satisfy Assumption A 5.2. However, those are only particular examples of approximating sequences satisfying Assumption A 5.2, and the algorithms that we will design will not use properties that are specific to these particular approximations.

- Also, note that we do not assume that \( f \) is twice continuously differentiable. Nevertheless, we may choose to use a second order local approximation \( \hat{f}_n \) of the function \( f \) in a neighborhood of \( x \), defined as follows:

\[
\hat{f}_n(y) := f_n(x) + \nabla_n f(x)^T (y - x) + \frac{1}{2} (y - x)^T \hat{\nabla}^2_n f(x) (y - x)
\]

We do not necessarily consider \( \hat{\nabla}^2_n f(x) \) as an approximation of \( \nabla^2 f(x) \), because \( \nabla^2 f(x) \) need not exist.

Assumptions A 5.1 and A 5.2 immediately lead to the following useful result.

Lemma 5.1. Suppose that Assumptions A 5.1 and A 5.2 hold. Then, for any sequence \( \{x_n\}_{n \in \mathbb{N}} \) in a compact subset of \( \mathcal{X} \), it holds that

\[
\lim_{n \to \infty} \left\| \nabla_n f(x_n) - \nabla f(x_n) \right\|_2 = 0
\]

Proof. We prove this result by contradiction. Consider any sequence \( \{x_n\}_{n \in \mathbb{N}} \) in a compact subset of \( \mathcal{X} \). Suppose that the assertion of Lemma 5.1 is false. Then, there exists \( \varepsilon > 0 \) and a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \), such that \( \left\| \nabla_{n_k} f(x_{n_k}) - \nabla f(x_{n_k}) \right\|_2 > \varepsilon \) for all \( k \in \mathbb{N} \). Since \( \{x_n\}_{n \in \mathbb{N}} \) is contained in a compact subset of \( \mathcal{X} \), \( \{x_{n_k}\}_{k \in \mathbb{N}} \) is contained in the same compact subset. Thus, there is a further subsequence \( \{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}} \) such that \( x_{m_k} \to x \) for some \( x \in \mathcal{X} \). It follows from Assumptions A 5.1 and A 5.2 that

\[
\lim_{k \to \infty} \left\| \nabla_{m_k} f(x_{m_k}) - \nabla f(x_{m_k}) \right\|_2 \leq \lim_{k \to \infty} \left\| \nabla_{m_k} f(x_{m_k}) - \nabla f(x) \right\|_2 + \lim_{k \to \infty} \left\| \nabla f(x) - \nabla f(x_{m_k}) \right\|_2 = 0
\]

However, since \( \{m_k\}_{k \in \mathbb{N}} \) is a subsequence of \( \{n_k\}_{k \in \mathbb{N}} \), it holds that \( \left\| \nabla_{m_k} f(x_{m_k}) - \nabla f(x_{m_k}) \right\|_2 > \varepsilon \) for all \( k \in \mathbb{N} \). This is a contradiction. Hence \( \left\| \nabla_n f(x_n) - \nabla f(x_n) \right\|_2 \to 0 \) as \( n \to \infty \).
For cases in which \( f \) and its first and possibly second order derivatives can be evaluated easily and accurately, a multitude of well-studied algorithms that require the exact evaluation of \( f(x) \), \( \nabla f(x) \), and possibly \( \nabla^2 f(x) \), at various \( x \in \mathcal{X} \), exist in the literature. We will refer to such algorithms as exact gradient algorithms. However, our aim is to develop inexact gradient algorithms that use only the approximating sequences \( \{f_n\}_{n \in \mathbb{N}} \) and \( \{\hat{\nabla}_n f\}_{n \in \mathbb{N}} \), and possibly a sequence \( \{\hat{\nabla}^2_n f\}_{n \in \mathbb{N}} \). Many of the inexact gradient algorithms that we study will be modifications of well-known exact gradient algorithms. Therefore, we first review some basic notions used in exact gradient algorithms. We start with a result regarding an optimality condition for \((P)\).

**Proposition 5.2.** Suppose that the feasible set \( \mathcal{X} \subset \mathbb{R}^l \) is convex, and the objective function \( f \) is continuously differentiable on a neighborhood of \( \mathcal{X} \).

(a) If \( x^* \) is a local minimum of \( f \) on \( \mathcal{X} \), then

\[
\nabla f(x^*)^T(x - x^*) \geq 0
\]  

for all \( x \in \mathcal{X} \).

(b) If \( f \) is also convex on \( \mathcal{X} \), then (108) is also sufficient for \( x^* \) to minimize \( f \) over \( \mathcal{X} \).

**Proof.** (a) Consider any local minimum \( x^* \) of \( f \) on \( \mathcal{X} \) and any \( x \in \mathcal{X} \). Then, there is \( \varepsilon > 0 \) such that \( f(x^*) \leq f(y) \) for all \( y \in \mathcal{X} \) with \( \|y - x^*\|_2 < \varepsilon \). Recall that the directional derivative satisfies

\[
\nabla f(x^*)^T(x - x^*) = \lim_{h \downarrow 0} \frac{f(x^* + h(x - x^*)) - f(x^*)}{h}
\]

Because \( f(x^*) \leq f(x^* + h(x - x^*)) \) for all \( h \in (0, \varepsilon / \|x - x^*\|_2) \), it follows that \( \nabla f(x^*)^T(x - x^*) \geq 0 \).

(b) Suppose that \( f \) is also convex on \( \mathcal{X} \). Then, for any \( x \in \mathcal{X} \),

\[
f(x) \geq f(x^*) + \nabla f(x^*)^T(x - x^*) \geq f(x^*)
\]

where the second inequality uses the result in part (a) that \( \nabla f(x^*)^T(x - x^*) \geq 0 \). Hence, (108) is also sufficient for optimality.

\[\square\]
Let
\[ S := \{ x^* \in \mathcal{X} : \nabla f(x^*)^T (x - x^*) \geq 0 \ \forall \ x \in \mathcal{X} \} \] (109)

Note that in general, (108) is only a necessary condition for \( x^* \) to be a local minimum of \( f \) on \( \mathcal{X} \), and therefore, the set of local minima of \( f \) on \( \mathcal{X} \) may be a strict subset of \( S \). We call any point \( x^* \in \mathcal{X} \) that satisfies (108) a *stationary point* of the function \( f \) on the set \( \mathcal{X} \).

Exact gradient algorithms are usually designed to search for stationary points of \( f \) in \( \mathcal{X} \). Typically, such algorithms follow the steps given below.

**Algorithm 5.1.**

**Step 1:** Choose the initial point \( x_0 \in \mathcal{X} \) and set the iteration counter \( n = 0 \).

**Step 2:** If \( x_n \in S \), then return the point \( x_n \) and stop.

**Step 3:** Find a point \( x_{n+1} \in \mathcal{X} \), increment the iteration counter to \( n + 1 \), and return to Step 2.

The first step is an initialization step that chooses the initial feasible solution. Then, at each iteration \( n \), Step 2 tests if the current solution \( x_n \) is a stationary point or not. For example, such a test can be performed by checking if
\[ \inf \{ \nabla f(x_n)^T (x - x_n) : x \in \mathcal{X} \} \geq 0 \] (110)

Step 2 is called a *stopping test* since the algorithm stops at iteration \( n \) if \( x_n \in S \). Otherwise, Step 3 finds a point \( x_{n+1} \in \mathcal{X} \) in finitely many operations and the algorithm returns to Step 2 to start the next iteration. Since we are interested in the local minima of \( f \), exact gradient algorithms are usually designed to be *descent algorithms*. That is, at each iteration \( n \), if \( x_n \not\in S \), then \( x_{n+1} \in \mathcal{X} \) is chosen such that \( f(x_{n+1}) < f(x_n) \).

Based on the steps of Algorithm 5.1, it is clear that only two outcomes are possible. Either the algorithm terminates at some iteration \( n \) and returns a stationary point \( x_n \), or it generates a sequence \( \{ x_n \}_{n \in \mathbb{N}} \subset \mathcal{X} \setminus S \). Consequently, the convergence analysis of a typical exact gradient algorithm includes the case when it fails to terminate. In particular, such analysis usually seeks to prove that any sequence \( \{ x_n \}_{n \in \mathbb{N}} \) generated by the algorithm converges to the set of stationary points.

Also, knowledge of \( f(x_n) \) and \( \nabla f(x_n) \) at each \( n \in \mathbb{N} \) is crucial for the execution of exact gradient algorithms. Indeed, Step 2 of Algorithm 5.1 requires knowledge of \( \nabla f(x_n) \) to perform the stopping
test. Also, in order for Algorithm 5.1 to be a descent algorithm, \( f(x_n) \) has to be evaluated exactly at every iteration \( n \).

Here, we consider inexact gradient algorithms that also search for points in \( S \) by executing a sequence of steps similar to those in Algorithm 5.1. A typical inexact gradient algorithm that we consider also starts at some \( x_0 \in \mathcal{X} \) and at each iteration \( n \), performs a stopping test. If the stopping test is successful, then the algorithm terminates and returns the current solution \( x_n \). Otherwise it finds \( x_{n+1} \in \mathcal{X} \) in finitely many operations and starts the next iteration. However, as mentioned earlier, inexact gradient algorithms only have access to the approximating functions \( f_n \), \( \hat{\nabla} f_n \), and possibly \( \hat{\nabla}^2 f_n \), at each iteration \( n \). Consequently, an inexact gradient algorithm cannot directly check if \( x_n \) is a stationary point of \( f \) on \( \mathcal{X} \) or not. Therefore, different stopping tests have to be designed for inexact gradient algorithms. For the same reason, inexact gradient algorithms also have to be designed to use \( f_n(x_n), \hat{\nabla} f_n(x_n), \) and possibly \( \hat{\nabla}^2 f_n(x_n) \) to generate the next point \( x_{n+1} \in \mathcal{X} \) at each iteration \( n \), instead of \( f(x_n), \nabla f(x_n), \) and possibly \( \nabla^2 f(x_n) \).

In this paper, we concern ourselves with the convergence analysis of such inexact gradient algorithms. Specifically, given any initial solution \( x_0 \in \mathcal{X} \), we wish to show that all the limit points of \( \{x_n\}_{n \in \mathbb{N}} \) lie in \( S \). Therefore, here we do not consider stopping tests for inexact gradient algorithms. First, we develop a framework for convergence analysis, and thereafter we use the framework for the convergence analysis of a variety of inexact gradient algorithms.

**Lemma 5.3.** Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathbb{R}^l \). Suppose that \( \{x_n\}_{n \in \mathbb{N}} \) is bounded. Let \( \mathcal{A} \) denote the set of limit points of \( \{x_n\}_{n \in \mathbb{N}} \). Then \( \mathcal{A} \) is nonempty and compact. In addition,

\[
\lim_{n \to \infty} d(x_n, \mathcal{A}) = 0
\]

(111)

**Proof.** Since \( \{x_n\}_{n \in \mathbb{N}} \) is bounded, it is contained in a compact subset of \( \mathbb{R}^l \). Thus \( \mathcal{A} \) is nonempty. Also, \( \mathcal{A} \) is closed, hence \( \mathcal{A} \) is compact.

We show by contradiction that \( \lim_{n \to \infty} d(x_n, \mathcal{A}) = 0 \). Suppose there is \( \varepsilon > 0 \) and a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( d(x_{n_k}, \mathcal{A}) > \varepsilon \) for all \( k \). Then, since \( \{x_{n_k}\}_{k \in \mathbb{N}} \) is bounded, it has a limit point \( x \in \mathcal{A} \). Also, since \( d(\cdot, \mathcal{A}) \) is continuous, it holds that \( d(x, \mathcal{A}) \geq \varepsilon \), which contradicts \( x \in \mathcal{A} \).
In this section we extend the framework of \( ? \) and \( ? \) for establishing convergence of descent algorithms. In this section, the objective function \( f : X \mapsto \mathbb{R} \) is required to be continuous on \( X \), but not necessarily differentiable. The set \( S \subset X \) of points at which \( f \) satisfies desired properties does not necessarily have to be the set of local minima of \( f \) on \( X \) or the set of stationary points of \( f \) on \( X \); however, the algorithms will be required to satisfy certain properties involving \( S \) specified below, and the stronger the choice of \( S \), the stronger the requirements that the algorithms have to satisfy and the stronger the conclusions of the results in this section. Let \( \text{cl}(S) \) denote the closure of \( S \). The next result shows that if the sequence \( \{x_n\}_{n \in \mathbb{N}} \) satisfies certain asymptotic descent properties, then asymptotically, the objective values \( f(x_n) \) are as good as the worst objective values in \( S \). Later we establish sufficient conditions for the asymptotic descent properties to hold, and we develop algorithms that satisfy these sufficient conditions.

**Theorem 6.1.** Consider a function \( f : X \mapsto \mathbb{R} \), a set \( S \subset X \), and a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset X \). Let \( A \) denote the set of accumulation points of \( \{x_n\}_{n \in \mathbb{N}} \) in \( X \). Suppose that \( \mu := \sup_{x \in S} f(x) < \infty \).

(If \( \sup_{x \in S} f(x) = \infty \), then the result follows immediately.) Suppose that the following assumptions hold:

**A 6.1.** The function \( f \) is continuous on \( X \).

**A 6.2.** There exists an accumulation point of \( \{x_n\}_{n \in \mathbb{N}} \) that lies in \( \text{cl}(S) \), i.e., \( A \cap \text{cl}(S) \neq \emptyset \).

**A 6.3.** For any \( \varepsilon > 0 \), there exists \( N_1(\varepsilon) \in \mathbb{N} \) such that \( f(x_{n+1}) \leq f(x_n) + \varepsilon \) for all \( n \geq N_1(\varepsilon) \).

**A 6.4.** For any \( \varepsilon > 0 \), there exists \( N_2(\varepsilon) \in \mathbb{N} \) such that for all \( n \geq N_2(\varepsilon) \), if \( \mu + \varepsilon \leq f(x_n) \leq \mu + 2\varepsilon \), then \( f(x_{n+1}) \leq f(x_n) \).

Then

\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in S} f(x) \tag{112}
\]

**Proof.** Consider any \( \varepsilon > 0 \). It follows from Assumption A 6.2 that there exists a subsequence of \( \{x_n\}_{n \in \mathbb{N}} \) that converges to a point in \( \text{cl}(S) \), and thus it follows from the continuity of \( f \) that \( f(x_n) < \mu + 2\varepsilon \) for infinitely many \( n \). Hence, there exists \( N \geq \max\{N_1(\varepsilon), N_2(\varepsilon)\} \) such that \( f(x_N) \leq \mu + 2\varepsilon \). If \( f(x_N) \leq \mu + \varepsilon \), then it follows from Assumption A 6.3 that \( f(x_{N+1}) \leq f(x_N) + \varepsilon \leq \mu + 2\varepsilon \); otherwise,
if $\mu + \varepsilon \leq f(x_N) \leq \mu + 2\varepsilon$, then it follows from Assumption A 6.4 that $f(x_{N+1}) \leq f(x_N) \leq \mu + 2\varepsilon$. Thus, by induction, $f(x_n) \leq \mu + 2\varepsilon$ for all $n \geq N$. Therefore, $\limsup_{n\to\infty} f(x_n) \leq \mu + 2\varepsilon$. Since $\varepsilon$ can be arbitrarily close to 0, it follows that $\limsup_{n\to\infty} f(x_n) \leq \mu$.

Next we show that, under some conditions, all the limit points of $\{x_n\}_{n\in\mathbb{N}}$ are in $\text{cl}(S)$.

**Theorem 6.2.** Consider a function $f : X \to \mathbb{R}$, a set $S \subset X$, and a sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$. Let $\mathcal{A}$ denote the set of accumulation points of $\{x_n\}_{n\in\mathbb{N}}$ in $X$. Suppose that Assumptions A 6.1 and A 6.3 of Theorem 6.1 hold. In addition, suppose that the following assumptions hold:

**A 6.5.** There exists $\bar{\mu} \in \mathbb{R}$ such that $\emptyset \neq \{x \in \mathcal{A} : f(x) \leq \bar{\mu}\} \subset \text{cl}(S)$.

**A 6.6.** For any $\varepsilon > 0$, there exists $N_2(\varepsilon) \in \mathbb{N}$ such that for all $n \geq N_2(\varepsilon)$, if $\bar{\mu} + \varepsilon \leq f(x_n) \leq \bar{\mu} + 2\varepsilon$, then $f(x_{n+1}) \leq f(x_n)$.

Then $\mathcal{A} \subset \text{cl}(S)$.

**Proof.** It follows in the same way as in the proof of Theorem 6.1 that $\limsup_{n\to\infty} f(x_n) \leq \bar{\mu}$. Thus, it follows from the continuity of $f$ that $f(x) \leq \bar{\mu}$ for any accumulation point $x$. Therefore $\mathcal{A} = \{x \in \mathcal{A} : f(x) \leq \bar{\mu}\} \subset \text{cl}(S)$.

**Corollary 6.3.** Consider a function $f : X \to \mathbb{R}$, a set $S \subset X$, and a sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$. Let $\mathcal{A}$ denote the set of accumulation points of $\{x_n\}_{n\in\mathbb{N}}$ in $X$. Suppose that Assumptions A 6.1 and A 6.3 of Theorem 6.1 hold. Let $\bar{\mu} := \inf_{x \in \mathcal{A}} f(x)$, and let $\tilde{S} := \arg\min_{x \in \mathcal{A}} f(x)$. Suppose that the following assumptions hold:

**A 6.7.** $\emptyset \neq \tilde{S} \subset \text{cl}(S)$.

**A 6.8.** For any $\varepsilon > 0$, there exists $N_2(\varepsilon) \in \mathbb{N}$ such that for all $n \geq N_2(\varepsilon)$, if $\bar{\mu} + \varepsilon \leq f(x_n) \leq \bar{\mu} + 2\varepsilon$, then $f(x_{n+1}) \leq f(x_n)$.

Then $\mathcal{A} = \tilde{S} \subset \text{cl}(S)$.

Let the distance between a point $x$ and a set $X \subset X$ be defined by $d(x, X) := \inf\{\|x - y\| : y \in X\}$, where $\|x - y\|$ is the distance between $x$ and $y$ in $X$. Next we show that, under some conditions, $d(x_n, S) \to 0$ as $n \to \infty$, that is, the whole sequence $\{x_n\}_{n\in\mathbb{N}}$ converges to $S$. 

67
Corollary 6.4. Suppose that the assumptions of Theorem 6.2 hold. In addition, suppose that the following assumption holds:

A 6.9. The sequence \( \{x_n\}_{n \in \mathbb{N}} \) remains in a compact subset of \( \mathcal{X} \).

Then \( d(x_n, \mathcal{S}) \to 0 \) as \( n \to \infty \).

Proof. We show the result by contradiction. Suppose that there is \( \varepsilon > 0 \) such that \( d(x_{n_k}, \mathcal{S}) > \varepsilon \) for a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \). Because \( \{x_{n_k}\}_{k \in \mathbb{N}} \) remains in a compact subset of \( \mathcal{X} \), there is a further subsequence \( \{j_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}} \) and \( x \in \mathcal{A} \) such that \( x_{j_k} \to x \) as \( k \to \infty \). It follows from the continuity of \( d \) that \( d(x, \mathcal{S}) \geq \varepsilon \). However, this contradicts the result of Theorem 6.2 that \( \mathcal{A} \subset \text{cl}(\mathcal{S}) \). \( \square \)

Next we provide sufficient conditions for the assumptions of Theorem 6.1 and Theorem 6.2 to be satisfied. Specifically, the asymptotic descent property provides a single criterion that, with continuity of \( f \) and compactness of \( \mathcal{X} \), is sufficient for Assumptions A6.2 – A6.5 and Assumption A 6.7 to hold.

Definition 6.1. Consider a function \( f : \mathcal{X} \mapsto \mathbb{R} \) and a set \( \mathcal{S} \subset \mathcal{X} \). Let \( \mu := \sup_{x \in \mathcal{S}} f(x) \). It is said that a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) satisfies the asymptotic descent property if for any subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( (x_{n_k}, x_{n_k+1}) \to (\tilde{x}, \bar{x}) \) for some \( \tilde{x}, \bar{x} \in \mathcal{X} \), it holds that

- if \( f(\tilde{x}) > \mu \), then \( f(\bar{x}) < f(\tilde{x}) \), and
- if \( f(\tilde{x}) \leq \mu \), then \( f(\bar{x}) \leq f(\tilde{x}) \).

It is said that an algorithm satisfies the asymptotic descent property if for any \( x_0 \in \mathcal{X} \), the sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by the algorithm satisfies the asymptotic descent property.

Note that the asymptotic descent property is only required to hold at the limiting pair \( (\tilde{x}, \bar{x}) \), i.e., the algorithm need not result in descent at every iteration. Many algorithms satisfy the following stronger form of the asymptotic descent property.

Lemma 6.5. Consider function \( f : \mathcal{X} \mapsto \mathbb{R} \), set \( \mathcal{S} \subset \mathcal{X} \), and a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \). Suppose that for any subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( (x_{n_k}, x_{n_k+1}) \to (\tilde{x}, \bar{x}) \) for some \( \tilde{x}, \bar{x} \in \mathcal{X} \), it holds that
• if $\tilde{x} \notin S$, then $f(\bar{x}) < f(\tilde{x})$, and
• if $\tilde{x} \in S$, then $f(\bar{x}) \leq f(\tilde{x})$.

Then the sequence $\{x_n\}_{n \in \mathbb{N}}$ satisfies the asymptotic descent property.

Next we show that if Assumptions A 6.1 and A 6.9 hold and the asymptotic descent property is satisfied, then Assumptions A6.2 – A6.5 and Assumption A 6.7 are satisfied.

**Lemma 6.6.** Suppose that Assumptions A 6.1 and A 6.9 hold. If a sequence $\{x_n\}_{n \in \mathbb{N}}$ satisfies the asymptotic descent property, then $\emptyset \neq \tilde{S} := \arg\min_{x \in A} f(x) \subset S$, and thus $\{x_n\}_{n \in \mathbb{N}}$ satisfies Assumptions A6.2, A6.5, and A6.7.

**Proof.** By Assumption A 6.9, $\{x_n\}_{n \in \mathbb{N}}$ remains in a compact subset of $X$, and thus $A$ is nonempty, and also $A$ is a closed subset of a compact set and hence is compact. By Assumption A 6.1, $f$ is continuous, and thus $\tilde{S} := \arg\min_{x \in A} f(x)$ is nonempty and compact. Consider any $\tilde{x} \in \arg\min_{x \in A} f(x)$. Let $\{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N}$ be a subsequence such that $x_{n_k} \to \tilde{x}$ as $k \to \infty$. Since $\{x_n\}_{n \in \mathbb{N}}$ remains in a compact subset of $X$, $\{x_{n_k+1}\}_{k \in \mathbb{N}}$ has a convergent subsequence. Therefore, there is a further subsequence $\{j_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}}$ such that $x_{j_k} \to \bar{x}$ and $x_{j_k+1} \to \bar{x} \in A$ as $k \to \infty$. Further, it follows from the asymptotic descent property that if $\tilde{x} \notin S$ then $f(\bar{x}) < f(\tilde{x})$, which contradicts $\tilde{x} \in \arg\min_{x \in A} f(x)$. Therefore $\tilde{x} \in S$ and thus $\arg\min_{x \in A} f(x) \subset S$. 

**Lemma 6.7.** Suppose that Assumptions A 6.1 and A 6.9 hold. If a sequence $\{x_n\}_{n \in \mathbb{N}}$ satisfies the asymptotic descent property, then $\{x_n\}_{n \in \mathbb{N}}$ satisfies Assumption A 6.3. That is, for any $\varepsilon > 0$, there exists $N_1(\varepsilon) \in \mathbb{N}$ such that $f(x_{n+1}) \leq f(x_n) + \varepsilon$ for all $n \geq N_1(\varepsilon)$.

**Proof.** We show the result by contradiction. Suppose that there exists an $\varepsilon > 0$ and a subsequence $\{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N}$ such that $f(x_{n_{k+1}}) > f(x_{n_k}) + \varepsilon$ for all $k \in \mathbb{N}$. By Assumption A 6.9, $\{x_n\}_{n \in \mathbb{N}}$ remains in a compact subset of $X$, and hence there is a further subsequence $\{j_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}}$ such that $x_{j_k} \to \bar{x}$ and $x_{j_k+1} \to \bar{x}$ as $k \to \infty$. It follows from the continuity of $f$ that $f(\bar{x}) \geq f(\bar{x}) + \varepsilon$. However, it follows from the asymptotic descent property that $f(\bar{x}) \leq f(\tilde{x})$, and thus a contradiction is reached.

**Lemma 6.8.** Suppose that Assumptions A 6.1 and A 6.9 hold. If a sequence $\{x_n\}_{n \in \mathbb{N}}$ satisfies the asymptotic descent property, then $\{x_n\}_{n \in \mathbb{N}}$ satisfies Assumption A 6.4. That is, for any $\varepsilon > 0$,
there exists \( N_2(\varepsilon) \in \mathbb{N} \) such that for all \( n \geq N_2(\varepsilon) \), if \( \mu + \varepsilon \leq f(x_n) \leq \mu + 2\varepsilon \), then \( f(x_{n+1}) \leq f(x_n) \).

**Proof.** We show the result by contradiction. Suppose that there exists \( \varepsilon > 0 \) and a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( \mu + \varepsilon \leq f(x_{n_k}) \leq \mu + 2\varepsilon \) and \( f(x_{n_k+1}) > f(x_{n_k}) \) for all \( k \in \mathbb{N} \). By Assumption A 6.9, \( \{x_n\}_{n \in \mathbb{N}} \) remains in a compact subset of \( \mathcal{X} \), and hence there is a further subsequence \( \{j_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}} \) such that \( x_{j_k} \to \tilde{x} \) and \( x_{j_k+1} \to \bar{x} \) as \( k \to \infty \). It follows from the continuity of \( f \) that \( \mu + \varepsilon \leq f(\tilde{x}) \leq \mu + 2\varepsilon \) and \( f(\bar{x}) \geq f(\tilde{x}) \). It follows from the asymptotic descent property that \( f(\bar{x}) < f(\tilde{x}) \), and thus a contradiction is reached. \( \square \)

It follows from Lemmas 6.6 through 6.8 that if Assumptions A 6.1 and A 6.9 are satisfied and sequence \( \{x_n\}_{n \in \mathbb{N}} \) satisfies the asymptotic descent property, then the conclusion of Theorem 6.1 (and of Corollary 6.4 if Assumptions A 6.5 and A 6.6 are also satisfied) holds.

**Theorem 6.9.** Consider a function \( f : \mathcal{X} \mapsto \mathbb{R} \), a set \( \mathcal{S} \subset \mathcal{X} \), and a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \). Suppose that Assumptions A 6.1 and A 6.9 are satisfied, and that \( \{x_n\}_{n \in \mathbb{N}} \) satisfies the asymptotic descent property. Then

\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in \mathcal{S}} f(x)
\]

In addition, if Assumptions A 6.5 and A 6.6 are satisfied, then

\[
\lim_{n \to \infty} d(x_n, \mathcal{S}) = 0
\]

Next, we develop algorithms that satisfy the sufficient conditions for asymptotic descent identified in Lemma 6.5.

### 7 Line Search Methods

In this section, we continue with the setup of Section 5, and we develop line search based algorithms for problem (P). We first provide a general description of the steps used by typical line search algorithms to generate the next point \( x_{n+1} \in \mathcal{X} \) at iteration \( n \), given \( \{x_0, \ldots, x_n\} \). Given a set \( \mathcal{X} \subset \mathbb{R}^l \) and \( \bar{\alpha} > 0 \), a function \( h : [0, \bar{\alpha}] \mapsto \mathcal{X} \) is called an arc in \( \mathcal{X} \). (If \( \bar{\alpha} = \infty \), then we consider a function \( h : [0, \infty) \mapsto \mathcal{X} \).)

Next we give a generic description of an iteration of a line search algorithm.
**Step 1:** Choose an arc $h_n$.

**Step 2:** Search the arc $\{h_n(\alpha) : \alpha \in [0, \bar{\alpha}]\}$ for an appropriate next point $x_{n+1} = h_n(\alpha_n)$ for some $\alpha_n \in [0, \bar{\alpha}]$.

The method used to choose the arc $h_n$ is called the arc rule. The chosen value $\alpha_n$ is called the step-size, and the method used to choose $\alpha_n$ is called the step-size rule.

As usual, the sequences $\{h_n\}_{n\in\mathbb{N}}$ and $\{\alpha_n\}_{n\in\mathbb{N}}$ may depend on the sequence $\{x_n\}_{n\in\mathbb{N}}$ of iterates, in the sense that each $h_n$ and $\alpha_n$ may depend on $x_0, \ldots, x_n$, as well as the function values, derivatives, and other data computed up to iteration $n$. However, this dependence is not shown in the notation.

Most line search algorithms use exact evaluations of both $f$ and $\nabla f$, and sometimes $\nabla^2 f$, in both the arc selection step and the arc search step described above. We will develop line search algorithms that use the approximations $f_n$, $\hat{\nabla} f_n$, and sometimes $\hat{\nabla}^2 f_n$, at each iteration $n$.

The rest of this section is structured as follows. First, we give examples of some commonly used arc rules and describe some of their useful properties. For each of these examples, we show how the arc rule can be modified to use $f_n$, $\hat{\nabla} f_n$, and sometimes $\hat{\nabla}^2 f_n$. Further, we establish some useful properties of such arc rules based on the approximations. Subsequently, we consider three widely used step-size rules. We show how each step-size rule can be modified to use $f_n$, $\hat{\nabla} f_n$, and sometimes $\hat{\nabla}^2 f_n$. Then we form line search algorithms by combining various arc rules and step-size rules that use the approximations, and we analyze the convergence of such algorithms by using the framework developed in Section 6.

We will require the following notation. Any $H \in S_{++}^{l \times l}$ defines a norm on $\mathbb{R}^l$ as follows:

$$\|x\|_H := \sqrt{x^T H x}$$

We will use such norms to define various arc rules. Also, given constants $0 < \lambda_1 < \lambda_2 < \infty$, set $\mathcal{P} \subset S_{++}^{l \times l}$ is defined by

$$\mathcal{P} := \{H \in S_{++}^{l \times l} : \lambda_1 \leq \lambda^\min(H) \leq \lambda^\max(H) \leq \lambda_2\} \quad (113)$$

?, pp. 539–540, shows that the roots of a polynomial are continuous in the coefficients of the polynomial, and thus the eigenvalues of a matrix are continuous in the entries of the matrix. It
follows that the finite dimensional set $P$ is closed and bounded, and hence compact. Note that
\[
\lambda_1 \|x\|_2^2 \leq \lambda_{\min}(H) \|x\|_2^2 \leq \|x\|_\mu^2 \leq \lambda_{\max}(H) \|x\|_2^2 \leq \lambda_2 \|x\|_2^2
\] (114)
for any $H \in P$ and $x \in \mathbb{R}^l$.

In order to show the convergence of our line search algorithms, we will consider sequences of points of the form $\{h_n(\alpha)\}_{n \in \mathbb{N}}$. Let $\mathcal{Y}(\alpha)$ denote the set of limit points of the sequence $\{h_n(\alpha)\}_{n \in \mathbb{N}}$ for each step-size $\alpha \in [0, \bar{\alpha}]$.

**Lemma 7.1.** Consider a sequence $\{h_n\}_{n \in \mathbb{N}}$ of arcs. Suppose that for any sequence $\{\alpha_n\}_{n \in \mathbb{N}} \subset (0, \bar{\alpha}]$ such that $\alpha_n \to 0$ as $n \to \infty$, it holds that
\[
\lim_{n \to \infty} h_n(\alpha_n) = x
\]
for some $x \in \mathbb{R}^l$. Suppose that $\mathcal{Y}(\alpha) \neq \emptyset$ for all $\alpha \in (0, \bar{\alpha}]$. Then
\[
\lim_{\alpha \downarrow 0} \sup_{y \in \mathcal{Y}(\alpha)} \|y - x\|_2 = 0
\]

**Proof.** We show the result by contradiction. Suppose that there is $\delta > 0$, a sequence $\{\alpha_n\}_{n \in \mathbb{N}} \subset (0, \bar{\alpha}]$ such that $\alpha_n \to 0$ as $n \to \infty$, and a sequence $\{y_n\}_{n \in \mathbb{N}}$ such that $y_n \in \mathcal{Y}(\alpha_n)$ and $\|y_n - x\|_2 > \delta$ for all $n \in \mathbb{N}$. Because $y_1 \in \mathcal{Y}(\alpha_1)$, there is a subsequence of $\{h_n(\alpha_1)\}_{n \in \mathbb{N}}$ that converges to $y_1$. Thus there is $n_1 \in \mathbb{N}$ such that $\|h_{n_1}(\alpha_1) - y_1\|_2 < \delta/2$, and hence $\|h_{n_1}(\alpha_1) - x\|_2 > \delta/2$. Similarly, there is $n_2 > n_1$ such that $\|h_{n_2}(\alpha_2) - y_2\|_2 < \delta/2$, and hence $\|h_{n_2}(\alpha_2) - x\|_2 > \delta/2$. Continuing by induction, for each $k \in \mathbb{N}$, there is $n_k > n_{k-1}$ such that $\|h_{n_k}(\alpha_k) - y_k\|_2 < \delta/2$, and hence $\|h_{n_k}(\alpha_k) - x\|_2 > \delta/2$. Construct the sequence $\{\alpha'_n\}_{n \in \mathbb{N}} \subset (0, \bar{\alpha}]$ by choosing $\alpha'_n = \alpha_1$ for $n = 1, \ldots, n_1$, and in general, $\alpha'_n = \alpha_k$ for $n = n_{k-1} + 1, \ldots, n_k$. Thus we have constructed a subsequence $\{h_{n_k}\}_{k \in \mathbb{N}} \subset \{h_n\}_{n \in \mathbb{N}}$ and a sequence $\{\alpha'_n\}_{n \in \mathbb{N}} \subset (0, \bar{\alpha}]$ such that $\alpha'_n \to 0$ as $n \to \infty$ and $\|h_{n_k}(\alpha'_n) - x\|_2 = \|h_{n_k}(\alpha_k) - x\|_2 > \delta/2$ for all $k \in \mathbb{N}$, which contradicts the assumption that for any sequence $\{\alpha_n\}_{n \in \mathbb{N}} \subset (0, \bar{\alpha}]$ such that $\alpha_n \to 0$ as $n \to \infty$, it holds that $\lim_{n \to \infty} h_n(\alpha_n) = x$. \qed

### 7.1 Arc Rules

In this section we discuss three examples of widely used arc rules. We make the following assumption throughout this section.

**A 7.1.** The feasible set $X \subset \mathbb{R}^l$ is nonempty, closed and convex.
7.1.1 Scaled Gradient Projection

In this section we outline the scaled gradient projection method of generating arcs. For any \( y \in \mathbb{R}^l \) and \( H \in \mathbb{S}_{++}^{l \times l} \), it can be shown that the problem

\[
\min_{x \in \mathcal{X}} \| y - x \|_H^2 \\
\text{s.t.} \quad x \in \mathcal{X}
\]

has a unique optimal solution. Thus, the projection function \( \Pi_{\mathcal{X}} : \mathbb{R}^l \times \mathbb{S}_{++}^{l \times l} \mapsto \mathcal{X} \) is defined by

\[
\Pi_{\mathcal{X}}(y, H) := \arg \min_{x \in \mathcal{X}} \| y - x \|_H^2
\]

We refer to (GP) as the projection problem of \( y \) onto \( \mathcal{X} \) in the norm defined by \( H \), and the optimal solution \( \Pi_{\mathcal{X}}(y, H) \) is called the projection of \( y \) onto \( \mathcal{X} \) in the norm defined by \( H \).

Some useful properties of the projection function are given below.

\textbf{P 7.1.} For any sequence \( \{(y_n, H_n)\}_{n \in \mathbb{N}} \subset \mathbb{R}^l \times \mathbb{S}_{++}^{l \times l} \) such that

\[
\lim_{n \to \infty} \| y_n - y \|_2 = 0 \quad \text{and} \quad \lim_{n \to \infty} \| H_n - H \|_2 = 0
\]

for some \( (y, H) \in \mathbb{R}^l \times \mathbb{S}_{++}^{l \times l} \), it holds that

\[
\lim_{n \to \infty} \Pi_{\mathcal{X}}(y_n, H_n) = \Pi_{\mathcal{X}}(y, H)
\]

That is, the function \( \Pi_{\mathcal{X}} \) is continuous with respect to both its arguments.

\textbf{P 7.2.} For any \( y \in \mathbb{R}^l \) and \( H \in \mathbb{S}_{++}^{l \times l} \), \( x \in \mathcal{X} \) satisfies \( x = \Pi_{\mathcal{X}}(y, H) \) if and only if

\[
(y - x)^T H (z - x) \leq 0 \quad \forall \quad z \in \mathcal{X}
\]

\textbf{P 7.3.} The projection function is non-expansive in its first argument, i.e., for any \( y, z \in \mathbb{R}^l \) and \( H \in \mathbb{S}_{++}^{l \times l} \),

\[
\| \Pi_{\mathcal{X}}(y, H) - \Pi_{\mathcal{X}}(z, H) \|_H \leq \| y - z \|_H
\]

For any \( x \in \mathcal{X} \), \( d \in \mathbb{R}^l \), and \( H \in \mathbb{S}_{++}^{l \times l} \), \( \{ \Pi_{\mathcal{X}}(x + \alpha d, H) : \alpha \in [0, \infty) \} \) is an arc in \( \mathcal{X} \). Next we state some properties of the corresponding arc rule \( h(\alpha) := \Pi_{\mathcal{X}}(x + \alpha d, H) \).

\textbf{P 7.4.} For any \( x \in \mathcal{X} \), \( d \in \mathbb{R}^l \), \( H \in \mathbb{S}_{++}^{l \times l} \), and \( \alpha \in [0, \infty) \),

\[
\| \Pi_{\mathcal{X}}(x + \alpha d, H) - x \|_H \leq \alpha \| d \|_H
\]
Also, for any $0 < \alpha_1 < \alpha_2 < \infty$, the following inequalities hold.

\[
\|\Pi_x(x + \alpha_1 d, H) - x\|_H \leq \|\Pi_x(x + \alpha_2 d, H) - x\|_H \leq \frac{\|\Pi_x(x + \alpha_1 d, H) - x\|_H}{\alpha_1} \geq \frac{\|\Pi_x(x + \alpha_2 d, H) - x\|_H}{\alpha_2}
\] (118)

\[
d^T H(\Pi_x(x + \alpha_1 d, H) - x) \leq d^T H(\Pi_x(x + \alpha_2 d, H) - x)
\] (119)

P 7.5. For any $x \in \mathcal{X}$, $d \in \mathbb{R}^l$, $H \in S^{l \times l}_{++}$, and $\alpha \in [0, \infty)$, it follows from Property P 7.2 that

\[
(x + \alpha d - \Pi_x(x + \alpha d, H))^T H (x - \Pi_x(x + \alpha d, H)) \leq 0
\]

Rearranging, it follows that for any $\alpha > 0$, $d^T H (x - \Pi_x(x + \alpha d, H)) \leq -\text{min}_H(H) \frac{\|x - \Pi_x(x + \alpha d, H)\|_2^2}{\alpha}$

(120)

where the last inequality follows from (114).

P 7.6. Suppose that $f$ is continuously differentiable on a neighborhood of $\mathcal{X}$, and let $S$ denote the set of stationary points of $f$ on $\mathcal{X}$, as defined in (109).

1. If $x^* \in \mathcal{X} \setminus S$, then

\[
\|\Pi_x(x^* - \alpha H^{-1}\nabla f(x^*), H) - x^*\|_2 > 0 \quad \text{for all } H \in S^{l \times l}_{++} \text{ and } \alpha > 0
\]

2. If $x^* \in S$, then

\[
x^* = \Pi_x(x^* - \alpha H^{-1}\nabla f(x^*), H) \quad \text{for all } H \in S^{l \times l}_{++} \text{ and } \alpha \geq 0
\]

P 7.7. Suppose that $f$ is continuously differentiable on a neighborhood of $\mathcal{X}$, and let $S$ denote the set of stationary points of $f$ on $\mathcal{X}$, as defined in (109). For any $x \in \mathcal{X} \setminus S$, $H \in S^{l \times l}_{++}$, and $\alpha > 0$, $\Pi_x(x - \alpha H^{-1}\nabla f(x), H) \in \mathcal{X}$, and thus, since $\mathcal{X}$ is convex, it follows that $\Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x$ is a feasible direction at $x$. That is, there is $\delta > 0$ such that $x + \delta[\Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x] \in \mathcal{X}$ for all $\delta \in [0, \delta]$. Also, using $d = -H^{-1}\nabla f(x)$ in Property P 7.5 and Property P 7.6 it follows that

\[
\nabla f(x)^T(\Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x) \leq -\frac{\text{min}_H(H)}{\alpha} \|\Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x\|_2^2 < 0
\]

(122)

for all $H \in S^{l \times l}_{++}$ and $\alpha > 0$. 

74
If \( \nabla f \) can be evaluated exactly, then a sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs corresponding to a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset X \) can be chosen as follows. At iteration \( n \), choose a matrix \( H_n \in \mathcal{P} \). For example, if \( f \) is twice continuously differentiable at \( x_n \), \( \nabla^2 f(x_n) \in \mathcal{P} \), and \( \nabla^2 f(x_n) \) can be evaluated exactly, then we can choose \( H_n = \nabla^2 f(x_n) \). The corresponding arc is given by

\[
h_n(\alpha) := \Pi_X(x_n - \alpha H_n^{-1} \nabla f(x_n), H_n)
\]

(123)

If \( \nabla f \) cannot be evaluated exactly, then we can choose

\[
h_n(\alpha) := \Pi_X(x_n - \alpha H_n^{-1} \hat{\nabla}_n f(x_n), H_n)
\]

(124)

Also, if \( \hat{\nabla}_n^2 f(x_n) \in \mathcal{P} \), we can choose \( H_n = \hat{\nabla}_n^2 f(x_n) \). Note that it follows from the continuity of \( \Pi_X \) that \( h_n \) is continuous, and \( x_n \in X \) implies that \( h_n(0) = x_n \).

Next, we point out some properties of sequences \( \{h_n\}_{n \in \mathbb{N}} \) of arcs given in (124) that are useful in line search algorithms.

**Lemma 7.2.** Consider an arc \( h_n \) defined in (124), with \( H_n \in \mathcal{P} \). Then

\[
\|h_n(\alpha) - x_n\|_2 \leq \frac{\alpha}{\lambda_1} \left\| \hat{\nabla}_n f(x_n) \right\|_2
\]

(125)

**Proof.** It follows from (117) in Property P 7.4 and (114) that

\[
\|h_n(\alpha) - x_n\|_2 = \left\| \Pi_X \left( x_n - \alpha H_n^{-1} \hat{\nabla}_n f(x_n), H_n \right) - x_n \right\|_2
\]

\[
\leq \frac{1}{\lambda_1} \left\| \Pi_X \left( x_n - \alpha H_n^{-1} \hat{\nabla}_n f(x_n), H_n \right) - x_n \right\|_{H_n}
\]

\[
\leq \frac{\alpha}{\sqrt{\lambda_1}} \left\| H_n^{-1} \hat{\nabla}_n f(x_n) \right\|_{H_n}
\]

\[
= \frac{\alpha}{\sqrt{\lambda_1}} \sqrt{\left( \hat{\nabla}_n f(x_n) \right)^T H_n^{-1} \hat{\nabla}_n f(x_n)}
\]

\[
\leq \frac{\alpha}{\lambda_1} \left\| \hat{\nabla}_n f(x_n) \right\|_2
\]

\[\square\]

**Lemma 7.3.** Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset X \) such that \( x_n \rightarrow x \in X \) as \( n \rightarrow \infty \). Consider a corresponding sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs defined in (124), with \( \{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P} \). Suppose that \( f \) is continuously differentiable on a neighborhood of \( X \), and that \( \{\hat{\nabla}_n f\}_{n \in \mathbb{N}} \) satisfies Assumption A 5.2.

Then, for any \( \bar{\alpha} \in [0, \infty) \) and any sequence \( \{\alpha_n\}_{n \in \mathbb{N}} \subset [0, \bar{\alpha}] \), the sequence \( \{h_n(\alpha_n)\}_{n \in \mathbb{N}} \) is bounded.

**Proof.**
It follows from Lemma 7.2 that

$$
\|h_n(\alpha_n) - x_n\|_2 \leq \frac{\alpha_n}{\lambda_1}\|\nabla f(x_n)\|_2
$$

Since \(\hat{\nabla} f(x_n) \to \nabla f(x)\), \(\{\hat{\nabla} f(x_n)\}_{n \in \mathbb{N}}\) is bounded. Also, \(\{\alpha_n\}_{n \in \mathbb{N}}\) is bounded. Thus \(\{\|h_n(\alpha_n) - x_n\|_2\}_{n \in \mathbb{N}}\) is bounded. But \(x_n \to x\), and therefore \(\{h_n(\alpha_n)\}_{n \in \mathbb{N}}\) is bounded. \(\square\)

Consider a sequence \(\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X}\) such that \(x_n \to x \in \mathcal{X}\) as \(n \to \infty\). Let \(\mathcal{H}\) denote the set of accumulation points of the corresponding sequence \(\{H_n\}_{n \in \mathbb{N}}\). (Thus \(\mathcal{H}\) depends on \(\{x_n\}_{n \in \mathbb{N}}\), but this dependence is not shown in the notation.) Since \(\{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P}\) and \(\mathcal{P}\) is compact, it follows that \(\mathcal{H}\) is a nonempty compact subset of \(\mathcal{P}\). Also, it follows from Theorem 2.3.4, that if \(H_n \to H\), then \(H_n^{-1} \to H^{-1}\).

**Lemma 7.4.** Consider a sequence \(\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X}\) such that \(x_n \to x \in \mathcal{X}\) as \(n \to \infty\). Consider a corresponding sequence \(\{h_n\}_{n \in \mathbb{N}}\) of arcs defined in (124), with \(\{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P}\). Suppose that \(f\) is continuously differentiable on a neighborhood of \(\mathcal{X}\), and that \(\{\hat{\nabla} f\}_{n \in \mathbb{N}}\) satisfies Assumption A 5.2. Then, for any \(\alpha \in [0, \infty)\),

$$
\mathcal{Y}(\alpha) = \left\{\Pi_x(x - \alpha H^{-1} \nabla f(x), H) : H \in \mathcal{H}\right\}
$$

**Proof.** Consider any \(H \in \mathcal{H}\). There exists a subsequence \(\{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N}\) such that \(H_{n_k} \to H\) as \(k \to \infty\). Also, \(x_{n_k} \to x\), and from Assumption A 5.2, \(\hat{\nabla} f(x_{n_k}) \to \nabla f(x)\). Thus, it follows from the continuity of \(\Pi_x\) (Property P 7.1) that

$$
\lim_{k \to \infty} h_{n_k}(\alpha) = \lim_{k \to \infty} \Pi_x(x_{n_k} - \alpha H_{n_k}^{-1} \hat{\nabla} f(x_{n_k}), H_{n_k}) = \Pi_x(x - \alpha H^{-1} \nabla f(x), H)
$$

Therefore \(\Pi_x(x - \alpha H^{-1} \nabla f(x), H) \in \mathcal{Y}(\alpha)\).

Consider any \(y \in \mathcal{Y}(\alpha)\). There exists a subsequence \(\{m_k\}_{k \in \mathbb{N}} \subset \mathbb{N}\) such that

$$
y = \lim_{k \to \infty} h_{m_k}(\alpha) = \lim_{k \to \infty} \Pi_x(x_{m_k} - \alpha H_{m_k}^{-1} \hat{\nabla} f(x_{m_k}), H_{m_k}) \quad (126)
$$

Since \(\{H_{m_k}\}_{k \in \mathbb{N}} \subset \mathcal{P}\), there is a further subsequence \(\{j_k\}_{k \in \mathbb{N}} \subset \{m_k\}_{k \in \mathbb{N}}\), and some \(H \in \mathcal{H}\), such that \(H_{j_k} \to H\) as \(k \to \infty\). Also, \(x_{j_k} \to x\), and from Assumption A 5.2, \(\hat{\nabla} f(x_{j_k}) \to \nabla f(x)\). Thus, it again follows from the continuity of \(\Pi_x\) (Property P 7.1) that

$$
y = \lim_{k \to \infty} \Pi_x(x_{j_k} - \alpha H_{j_k}^{-1} \hat{\nabla} f(x_{j_k}), H_{j_k}) = \Pi_x(x - \alpha H^{-1} \nabla f(x), H)
$$

Therefore, \(y \in \left\{\Pi_x(x - \alpha H^{-1} \nabla f(x), H) : H \in \mathcal{H}\right\}\). \(\square\)
Lemma 7.5. Consider a sequence \( \{ x_n \}_{n \in \mathbb{N}} \subset X \) such that \( x_n \to x \in X \) as \( n \to \infty \). Consider a corresponding sequence \( \{ h_n \}_{n \in \mathbb{N}} \) of arcs defined in (124), with \( \{ H_n \}_{n \in \mathbb{N}} \subset P \). Suppose that \( f \) is continuously differentiable on a neighborhood of \( X \), and that \( \{ \nabla_n f \}_{n \in \mathbb{N}} \) satisfies Assumption A 5.2. Let \( S \) denote the set of stationary points of \( f \) on \( X \), as defined in (109). Then \( \{ h_n \}_{n \in \mathbb{N}} \) satisfies the following properties.

**P 7.8.** For any \( \alpha \in [0, \infty) \), the sequence \( \{ h_n(\alpha) \}_{n \in \mathbb{N}} \) is bounded, and \( \mathcal{Y}(\alpha) \) is nonempty and compact.

**P 7.9.** For any sequence \( \{ \alpha_n \}_{n \in \mathbb{N}} \subset [0, \infty) \) such that \( \alpha_n \to 0 \) as \( n \to \infty \), it holds that
\[
\lim_{n \to \infty} h_n(\alpha_n) = x
\]

**P 7.10.**
\[
\lim_{\alpha \to 0} \sup_{y \in \mathcal{Y}(\alpha)} \| y - x \|_2 = 0
\]

**P 7.11.** If \( x \notin S \), then for any \( \alpha \in (0, \infty) \) and any \( y \in \mathcal{Y}(\alpha) \),
\[
\| y - x \|_2 > 0 \quad \text{and} \quad \nabla f(x)^T(y - x) \leq -\frac{\lambda_1}{\alpha} \| y - x \|_2^2 < 0
\]
Consequently, for any \( \bar{\alpha} \in (0, \infty) \),
\[
\sup_{\alpha \in (0, \bar{\alpha}]} \sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^T \frac{y - x}{\| y - x \|_2} < 0
\]

**P 7.12.** If \( x \in S \), then \( \mathcal{Y}(\alpha) = \{ x \} \) for every \( \alpha \in [0, \infty) \). Consequently, \( \nabla f(x)^T(y - x) = 0 \) for all \( \alpha \in [0, \infty) \) and \( y \in \mathcal{Y}(\alpha) \).

**Proof.** (a) **Property P 7.8:** It follows from Lemma 7.3 that for any \( \alpha \in [0, \infty) \), the sequence \( \{ h_n(\alpha) \}_{n \in \mathbb{N}} \) is bounded. Hence \( \mathcal{Y}(\alpha) \) is nonempty and compact.

(b) **Property P 7.9:** It follows from Lemma 7.2 that
\[
\| h_n(\alpha_n) - x_n \|_2 \leq \frac{\alpha_n}{\lambda_1} \| \hat{\nabla}_n f(x_n) \|_2
\]
Since \( \alpha_n \to 0 \) and \( \hat{\nabla}_n f(x_n) \to \nabla f(x) \), it follows that
\[
\lim_{n \to \infty} \| h_n(\alpha_n) - x_n \|_2 = 0
\]
Also, \( x_n \to x \), and thus \( h_n(\alpha_n) \to x \).
(c) **Property P 7.10:** Property P 7.10 follows from Property P 7.8, Property P 7.9, and Lemma 7.1.

(d) **Property P 7.11:** Consider any $\alpha \in (0, \infty)$ and any $y \in \mathcal{Y}(\alpha)$. It follows from Lemma 7.4 that there exists $H \in \mathcal{H}$ such that $y = \Pi_x (x - \alpha H^{-1} \nabla f(x), H)$. It follows from (refeqn:proj dir) of Property P 7.5 that

$$\nabla f(x)^T (y - x) \leq - \frac{\lambda_{\min}(H)}{\alpha} \|y - x\|_2^2 \leq - \frac{\lambda_1}{\alpha} \|y - x\|_2^2 \quad (127)$$

Next we show by contradiction that $\|y - x\|_2 > 0$. Suppose that $x = y \in \mathcal{Y}(\alpha)$. It follows from Lemma 7.4 that there exists $H \in \mathcal{H}$ such that $x = \Pi_x (x - \alpha H^{-1} \nabla f(x), H)$. It follows from Property P 7.6 that $x \in S$, which is a contradiction. Therefore $\|y - x\|_2 > 0$.

Again, using (refeqn:proj dir) of Property P 7.5, it follows that

$$\nabla f(x)^T (y - x) \leq - \frac{\|y - x\|_H^2}{\alpha} \leq - \frac{\sqrt{\lambda_1}}{\alpha} \|y - x\|_H \|y - x\|_H$$

Since $\|y - x\|_2 > 0$, it follows that

$$\nabla f(x)^T \frac{y - x}{\|y - x\|_2} \leq - \frac{\sqrt{\lambda_1}}{\alpha} \|y - x\|_H$$

Since $y = \Pi_x (x - \alpha H^{-1} \nabla f(x), H)$, it follows from (119) of Property P 7.4 that for all $\alpha \in (0, \bar{\alpha}]$,

$$\nabla f(x)^T \frac{y - x}{\|y - x\|_2} \leq - \frac{\sqrt{\lambda_1}}{\alpha} \|\Pi_x (x - \bar{\alpha} H^{-1} \nabla f(x), H) - x\|_H$$

$$\leq - \frac{\sqrt{\lambda_1}}{\alpha} \inf_{H \in \mathcal{H}} \|\Pi_x (x - \bar{\alpha} H^{-1} \nabla f(x), H) - x\|_H$$

Since the right side of the above inequality is independent of $\alpha$ and $y$, it follows that

$$\sup_{\alpha \in (0, \bar{\alpha}] \sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^T \frac{y - x}{\|y - x\|_2} \leq - \frac{\sqrt{\lambda_1}}{\alpha} \inf_{H \in \mathcal{H}} \|\Pi_x (x - \bar{\alpha} H^{-1} \nabla f(x), H) - x\|_H$$

Since $x \notin S$, it follows from Property P 7.6 that $\|\Pi_x (x - \bar{\alpha} H^{-1} \nabla f(x), H) - x\|_H > 0$ for all $H \in \mathcal{H}$. Since $\mathcal{H}$ is nonempty and compact, and $\|\Pi_x (x - \bar{\alpha} H^{-1} \nabla f(x), H) - x\|_H$ is a continuous function of $H$, it follows that $\inf_{H \in \mathcal{H}} \|\Pi_x (x - \bar{\alpha} H^{-1} \nabla f(x), H) - x\|_H > 0$. Therefore,

$$\sup_{\alpha \in (0, \bar{\alpha}] \sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^T \frac{y - x}{\|y - x\|_2} < 0$$

(e) **Property P 7.12:** Consider any $\alpha \in [0, \infty)$ and any $y \in \mathcal{Y}(\alpha)$. It follows from Lemma 7.4 that there exists $H \in \mathcal{H}$ such that $y = \Pi_x (x - \alpha H^{-1} \nabla f(x), H)$. Since $x \in S$, it follows from Property P 7.6 that $\Pi_x (x - \alpha H^{-1} \nabla f(x), H) = x$ for all $H \in S_{\alpha}^{\pi_f}$, and therefore $y = x$. 

$\square$
Lemma 7.6. Consider a sequence \( \{\alpha_n\}_{n \in \mathbb{N}} \subset (0, \bar{\alpha}] \), and a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that \( x_n \to x \in \mathcal{X} \) as \( n \to \infty \). Consider a corresponding sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs defined in (124), with \( \{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P} \). Suppose that \( f \) is continuously differentiable on a neighborhood of \( \mathcal{X} \), and that \( \{\hat{\nabla}_nf\}_{n \in \mathbb{N}} \) satisfies Assumption A.5.2. Let \( S \) denote the set of stationary points of \( f \) on \( \mathcal{X} \), as defined in (109). If
\[
\lim_{n \to \infty} \frac{\|h_n(\alpha_n) - x_n\|_2}{\alpha_n} = 0
\]
then \( x \in S \).

Proof. It follows from (114) and (119) of Property P 7.4 that
\[
\sqrt{\lambda_2} \frac{\|h_n(\alpha_n) - x_n\|_2}{\alpha_n} \geq \frac{\|h_n(\alpha_n) - x_n\|_{H_n}}{\alpha_n} \geq \frac{\|\Pi_x(x_n - \alpha_n H_n^{-1} \hat{\nabla}_n f(x_n), H_n) - x_n\|_{H_n}}{\alpha_n} \geq \frac{\|\Pi_x(x_n - \bar{\alpha} H_n^{-1} \hat{\nabla}_n f(x_n), H_n) - x_n\|_{H_n}}{\bar{\alpha}} \geq \sqrt{\lambda_1} \frac{\|\Pi_x(x_n - \bar{\alpha} H_n^{-1} \hat{\nabla}_n f(x_n), H_n) - x_n\|_2}{\bar{\alpha}}
\]
Since \( \lim_{n \to \infty} \frac{\|h_n(\alpha_n) - x_n\|_2}{\alpha_n} = 0 \), it follows that
\[
\lim_{n \to \infty} \frac{\|\Pi_x(x_n - \bar{\alpha} H_n^{-1} \hat{\nabla}_n f(x_n), H_n) - x_n\|_2}{\bar{\alpha}} = 0
\]
Since \( \{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P} \), there exists a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( H_{n_k} \to H \in \mathcal{H} \) as \( k \to \infty \). It follows from the continuity of \( \Pi_x \) that
\[
\lim_{k \to \infty} \frac{\|\Pi_x(x_{n_k} - \bar{\alpha} H_{n_k}^{-1} \hat{\nabla}_n f(x_{n_k}), H_{n_k}) - x_{n_k}\|_2}{\bar{\alpha}} = \frac{\|\Pi_x(x - \bar{\alpha} H^{-1} \nabla f(x), H) - x\|_2}{\bar{\alpha}} = 0
\]
Then it follows from Property P 7.6 that \( x \in S \).

7.1.2 Inexact Scaled Gradient Projection

In this section, we describe a variation of the gradient projection technique where we allow the projection to be calculated inexactly. Although gradient projection is an elegant way to obtain arcs in the feasible set at each iteration, typically it is expensive and unnecessary to compute the projections exactly. For example, when the constraints are affine, it is necessary to solve a quadratic
program to generate each point in the arc. We consider an approach that allows inaccuracy in the
calculation of the projection at every step and that retains desired convergence properties.

**Definition 7.1.** Let the constants \( \bar{\alpha}' \), \( \eta_1 \), and \( \zeta \) satisfy

\[
\bar{\alpha}' > 0, \quad \eta_1 \in (0, 1) \quad \text{and} \quad \zeta > 2
\]  

(128)

Consider any \( x \in \mathcal{X} \), \( \alpha \in [0, \bar{\alpha}'] \), \( d \in \mathbb{R}^l \), and \( H \in S^{l \times l}_{++} \). Then, a point \( y \in \mathcal{X} \) is called an *inexact projection* of the point \((x + \alpha d) \in \mathbb{R}^l \) onto the set \( \mathcal{X} \) in the norm defined by \( H \) if it satisfies

\[
\frac{3\alpha}{\sqrt{1 - \eta_1}} \|y - \Pi_{\mathcal{X}}(x + \alpha d, H)\|_H \|d\|_H \leq \left( \frac{\alpha}{\bar{\alpha}'} \right)^{\zeta/2 - 1} \|y - x\|_H^2 \quad \text{and} \quad (129)
\]

\[
\|y - \Pi_{\mathcal{X}}(x + \alpha d, H)\|_H \leq \frac{3\alpha}{\sqrt{1 - \eta_1}} \left( \frac{\alpha}{\bar{\alpha}'} \right)^{\zeta/2 - 1} \|d\|_H \quad (130)
\]

Let \( \hat{\Pi}_{\mathcal{X}}(x, \alpha, d, H) \) denote the set of inexact projections of \( x + \alpha d \) onto \( \mathcal{X} \).

Remarks:

- The exact projection \( \Pi_{\mathcal{X}}(x + \alpha d, H) \) satisfies both (129) and (130), and hence the set \( \hat{\Pi}_{\mathcal{X}}(x, \alpha, d, H) \) is nonempty. It is easy to verify that \( \hat{\Pi}_{\mathcal{X}}(x, \alpha, d, H) \) is a closed set and (130) establishes that \( \hat{\Pi}_{\mathcal{X}}(x, \alpha, d, H) \) is also bounded. Thus, \( \hat{\Pi}_{\mathcal{X}}(x, \alpha, d, H) \) is nonempty and compact for each \( x \in \mathcal{X} \), \( \alpha \in [0, \bar{\alpha}'] \), \( d \in \mathbb{R}^l \), and \( H \in S^{l \times l}_{++} \).

- The above definition of inexact projection is similar to the inexact projection idea described
in ?. It has been adapted here for use with projections in non-Euclidean norms.

Next we establish some useful properties of inexact projections before we describe their use in
optimization algorithms.

**Lemma 7.7.** Consider any \( x \in \mathcal{X} \), \( \alpha \in (0, \bar{\alpha}'] \), \( d \in \mathbb{R}^l \), \( H \in S^{l \times l}_{++} \), and \( y \in \hat{\Pi}_{\mathcal{X}}(x, \alpha, d, H) \). If

\[
0 < \frac{\alpha}{\bar{\alpha}'} \leq \left( \frac{1 - \eta_1}{4} \right)^{-1/2}
\]

then

\[
-d^T H (x - y) \geq \frac{\eta_1}{\alpha} \|x - y\|_H^2
\]

**Proof.** Recall the Cauchy-Schwarz inequality for the inner-product and associated norm defined on \( \mathbb{R}^l \) by \( H \):

\[
|x^T H y| \leq \|x\|_H \|y\|_H \quad \forall \ x, y \in \mathbb{R}^l \quad (131)
\]
It follows from Property P 7.5 and the Cauchy-Schwarz inequality that

\[-d^TH(x-y) = -d^TH(x - \Pi_x(x + \alpha d, H)) - d^TH(\Pi_x(x + \alpha d, H) - y)\]

\[\geq \frac{\|x - \Pi_x(x + \alpha d, H)\|^2}{\alpha} - \left|d^TH(\Pi_x(x + \alpha d, H) - y)\right|\]  \hspace{1cm} (132)

\[\geq \frac{\|x - \Pi_x(x + \alpha d, H)\|^2}{\alpha} - \|d\|_H \|\Pi_x(x + \alpha d, H) - y\|_H\]  \hspace{1cm} (133)

Consider the first term on the right side of (133).

\[\|x - \Pi_x(x + \alpha d, H)\|^2_H = \{[x - y] + [y - \Pi_x(x + \alpha d, H)]\}^T H \{[x - y] + [y - \Pi_x(x + \alpha d, H)]\} \]

\[= [x - y]^T H [x - y] + 2[y - \Pi_x(x + \alpha d, H)]^T H [x - y]\]

\[+ [y - \Pi_x(x + \alpha d, H)]^T H [y - \Pi_x(x + \alpha d, H)]\]

\[\geq \|x - y\|^2_H - 2 \|y - \Pi_x(x + \alpha d, H)\|^2_H\]  \hspace{1cm} (134)

It follows from the triangle inequality and (117) of Property P 7.4 that

\[-\|x - y\|_H \geq -\|x - \Pi_x(x + \alpha d, H)\|_H - \|\Pi_x(x + \alpha d, H) - y\|_H\]

\[\geq -\alpha \|d\|_H - \|y - \Pi_x(x + \alpha d, H)\|_H\]

Substituting the above inequality into (134), it follows that

\[\|x - \Pi_x(x + \alpha d, H)\|^2_H \geq \|x - y\|^2_H - 2\alpha \|y - \Pi_x(x + \alpha d, H)\|_H \|d\|_H - 2 \|y - \Pi_x(x + \alpha d, H)\|^2_H\]

Substituting the above inequality into (133), it follows that

\[-d^TH(x - y) \geq \frac{\|x - y\|^2_H}{\alpha} - 3 \|y - \Pi_x(x + \alpha d, H)\|_H \|d\|_H - 2 \frac{\|y - \Pi_x(x + \alpha d, H)\|^2}{\alpha}\]  \hspace{1cm} (135)

Since \(y \in \Pi_x(x, \alpha, d, H)\), it follows from applying (130) to the last term on the right side of (135) that

\[-d^TH(x - y) \geq \frac{\|x - y\|^2_H}{\alpha} - \frac{3}{\sqrt{1 - \eta_1}} \left[2 \left(\frac{\alpha}{\alpha'}\right)^{\zeta/2-1} + \sqrt{1 - \eta_1}\right] \|y - \Pi_x(x + \alpha d, H)\|_H \|d\|_H\]

Then it follows from (129) that

\[-d^TH(x - y) \geq \left[1 - 2 \left(\frac{\alpha}{\alpha'}\right)^{\zeta-2} - \sqrt{1 - \eta_1} \left(\frac{\alpha}{\alpha'}\right)^{\zeta/2-1}\right] \frac{\|x - y\|^2}{\alpha}\]
Thus, if
\[
\frac{\alpha}{\alpha'} \leq \left( \frac{1 - \eta_1}{4} \right)^{\frac{1}{\zeta - 2}}
\]
then
\[
-d^T H(x - y) \geq \frac{\eta}{\alpha} \|x - y\|^2_H
\]

**Corollary 7.8.** Suppose that $\bar{\alpha}$ satisfies
\[
0 < \bar{\alpha} \leq \left( \frac{1 - \eta_1}{4} \right)^{\frac{1}{\zeta - 2}} \bar{\alpha}'
\]
Then
\[
\nabla f(x)^T (y - x) \leq -\frac{\eta}{\alpha} \|y - x\|^2_H
\]
for all $x \in \mathcal{X}$, $\alpha \in (0, \bar{\alpha}]$, $H \in \mathbb{S}^{l \times l}_{++}$, and $y \in \Pi_{\mathcal{X}}(x, \alpha, -H^{-1}\nabla f(x), H)$.

**Corollary 7.9.** Suppose that $\bar{\alpha}$ satisfies
\[
0 < \bar{\alpha} \leq \left( \frac{1 - \eta_1}{4} \right)^{\frac{1}{\zeta - 2}} \bar{\alpha}'
\]
Then
\[
\hat{\nabla}_n f(x)^T (y - x) \leq -\frac{\eta}{\alpha} \lambda_1 \|y - x\|^2_2
\]
for all $x \in \mathcal{X}$, $\alpha \in (0, \bar{\alpha}]$, $H \in \mathcal{P}$, and $y \in \Pi_{\mathcal{X}}(x, \alpha, -H^{-1}\hat{\nabla}_n f(x), H)$.

**Lemma 7.10.** For any $x \in \mathcal{X}$, $\alpha \in [0, \bar{\alpha}'$, $H \in \mathbb{S}^{l \times l}_{++}$, and $y \in \Pi_{\mathcal{X}}(x, \alpha, -H^{-1}\nabla f(x), H)$,
\[
\|y - x\|_H \geq \frac{1}{2} \left\|\Pi_{\mathcal{X}}(x - \alpha H^{-1}\nabla f(x), H) - x\|_H \right\|
\]
(136)

Consequently, the following hold.

(a) If $x \notin \mathcal{S}$, then for any $\alpha \in (0, \bar{\alpha}']$ and $H \in \mathbb{S}^{l \times l}_{++}$,
\[
\inf_{y \in \Pi_{\mathcal{X}}(x, \alpha, -H^{-1}\nabla f(x), H)} \|y - x\|_H \geq \frac{1}{2} \left\|\Pi_{\mathcal{X}}(x - \alpha H^{-1}\nabla f(x), H) - x\|_H \right\| > 0
\]
(137)

Further, for any $H \in \mathbb{S}^{l \times l}_{++}$,
\[
\inf_{\alpha \in [0, \bar{\alpha}']} \inf_{y \in \Pi_{\mathcal{X}}(x, \alpha, -H^{-1}\nabla f(x), H)} \|y - x\|_H \geq \frac{1}{2} \inf_{\alpha \in (0, \bar{\alpha}']} \frac{\left\|\Pi_{\mathcal{X}}(x - \alpha H^{-1}\nabla f(x), H) - x\|_H \right\|}{\alpha} > 0
\]
(138)
(b) Otherwise, if \( x \in S \), then for all \( \alpha \in [0, \bar{\alpha}') \) and \( H \in S_{++}^{|n|} \),
\[
\hat{\Pi}_x(x, \alpha, -H^{-1}\nabla f(x), H) = \{ x \}
\]

**Proof.** It follows from (129) that
\[
\left( \frac{\alpha}{\bar{\alpha}'} \right)^{\zeta/2 - 1} \| y - x \|_H^2 \geq \frac{3\alpha}{\sqrt{1 - \eta_1}} \left( \frac{\alpha}{\bar{\alpha}'} \right)^{\zeta/2 - 1} \| y - \Pi_x(x - \alpha H^{-1}\nabla f(x), H) \|_H \| H^{-1}\nabla f(x) \|_H
\]
Also, it follows from (130) that
\[
\frac{3\alpha}{\sqrt{1 - \eta_1}} \left( \frac{\alpha}{\bar{\alpha}'} \right)^{\zeta/2 - 1} \| H^{-1}\nabla f(x) \|_H \geq \| y - \Pi_x(x - \alpha H^{-1}\nabla f(x), H) \|_H
\]
Combining the two inequalities above, it follows that
\[
\left( \frac{\alpha}{\bar{\alpha}'} \right)^{\zeta/2 - 1} \| y - x \|_H^2 \geq \| y - \Pi_x(x - \alpha H^{-1}\nabla f(x), H) \|_H^2 \tag{139}
\]
Also,
\[
\| x - \Pi_x(x - \alpha H^{-1}\nabla f(x), H) \|_H \leq \| x - y \|_H + \| y - \Pi_x(x - \alpha H^{-1}\nabla f(x), H) \|_H
\]
Thus it follows from (139) that
\[
\left[ 1 + \left( \frac{\alpha}{\bar{\alpha}'} \right)^{\zeta/2 - 1} \right] \| y - x \|_H \geq \| x - \Pi_x(x - \alpha H^{-1}\nabla f(x), H) \|_H
\]
Since \( \alpha \in [0, \bar{\alpha}'] \) and \( \zeta/2 - 1 > 0 \), it follows that
\[
\| y - x \|_H \geq \frac{1}{2} \| x - \Pi_x(x - \alpha H^{-1}\nabla f(x), H) \|_H
\]

(a) It follows from (136) that
\[
\inf_{y \in \Pi_x(x, \alpha, -H^{-1}\nabla f(x), H)} \| y - x \|_H \geq \frac{1}{2} \| \Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x \|_H
\]
Also, since \( x \notin S \), it follows from Property P 7.6 that \( \| \Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x \|_H > 0 \).
Therefore, (137) holds.

It also follows from (136) and from (119) of Property P 7.4 that
\[
\inf_{\alpha \in [0, \bar{\alpha}']} \inf_{y \in \Pi_x(x, \alpha, -H^{-1}\nabla f(x), H)} \| y - x \|_H \geq \frac{1}{2} \alpha \inf_{\alpha \in [0, \bar{\alpha}']} \| \Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x \|_H
\]
Also, since \( x \notin S \), it follows from Property P 7.6 that \( \| \Pi_x(x - \bar{\alpha}' H^{-1}\nabla f(x), H) - x \|_H > 0 \).
Therefore, (138) holds.
(b) Recall that since \( x \in S \), it follows from Property P 7.6 that \( x = \Pi_x(x - \alpha H^{-1} \nabla f(x), H) \).

Consider the following two cases.

Case(i): Suppose that \( \alpha \left\| H^{-1} \nabla f(x) \right\|_H > 0 \). Let

\[
\delta \coloneqq \left\| y - \Pi_x(x - \alpha H^{-1} \nabla f(x), H) \right\|_H = \left\| y - x \right\|_H
\]

Then it follows from (129) that

\[
3\alpha \left\| H^{-1} \nabla f(x) \right\|_H \delta \leq \sqrt{1 - \eta_1 \left( \frac{\alpha}{\alpha'} \right)^{\frac{\zeta}{2} - 2}} \delta^2
\]

Next it follows from (130) that

\[
\alpha \left\| H^{-1} \nabla f(x) \right\|_H \delta \leq \left( \frac{\alpha}{\alpha'} \right)^{\frac{\zeta}{2} - 2} \alpha \left\| H^{-1} \nabla f(x) \right\|_H \delta
\]

Since \( \alpha \left\| H^{-1} \nabla f(x) \right\|_H > 0 \), it follows that

\[
\delta \left[ 1 - \left( \frac{\alpha}{\alpha'} \right)^{\frac{\zeta}{2} - 2} \right] \leq 0
\]

Since \( \delta \geq 0 \), \( \alpha < \bar{\alpha}' \), and \( \zeta > 2 \), it follows that \( \delta = 0 \), that is, \( y = x \).

\[
\square
\]

Next, we discuss how the inexact projections can be used to generate an arc at each iteration of a line search algorithm. Suppose that the step-size rule that is used in the algorithm searches for a suitable step in the range \([0, \bar{\alpha}]\) for some \( \bar{\alpha} > 0 \). First, we choose constants \( \bar{\alpha}', \eta_1 \in (0, 1) \), and \( \zeta > 2 \), such that

\[
\bar{\alpha}' \geq \left( \frac{4}{1 - \eta_1} \right)^{\frac{1}{\zeta - 2}} \bar{\alpha}
\]

(140)

For any \( n \in \mathbb{N} \) and \( x_n \in X \), we choose \( H_n \in P \) as in Section 7.1.1. If \( \nabla f(x_n) \) can be evaluated exactly, then we choose an arc \( h_n \) such that

\[
h_n(\alpha) \in \hat{\Pi}_X(x_n, \alpha, -H_n^{-1} \nabla f(x_n), H_n) \quad \text{for each} \quad \alpha \in [0, \bar{\alpha}]
\]

(141)
If $f$ is twice continuously differentiable on a neighborhood of $\mathcal{X}$ and $\nabla^2 f(x_n)$ can be evaluated exactly, then we can choose $H_n = \nabla^2 f(x_n)$ if $\nabla^2 f(x_n) \in \mathcal{P}$. If we do not evaluate $\nabla f(x_n)$ exactly, then we use $\hat{\nabla}_n f(x_n)$, and choose an arc $h_n$ such that

$$h_n(\alpha) \in \tilde{\Pi}_x(x_n, \alpha, -H_n^{-1}\hat{\nabla}_n f(x_n), H_n) \quad \text{for each } \alpha \in [0, \tilde{\alpha}]$$  \hspace{1cm} (142)

Note that it follows from (130) that when $\alpha = 0$, $\tilde{\Pi}_x(x, 0, d, H) = \{\Pi_x(x, H)\}$ = \{x\} for each $x \in \mathcal{X}$, $d \in \mathbb{R}^l$, and $H \in \mathbb{S}_{++}^{l \times l}$.

Next we consider some properties of arc rules that satisfy (142).

**Lemma 7.11.** Consider an arc $h_n$ defined in (142), with $H_n \in \mathcal{P}$. Then

$$\|h_n(\alpha) - x_n\|_2 \leq \frac{\alpha}{\lambda_1} \left[1 + \frac{3}{\sqrt{1 - \eta_1}} \left(\frac{\alpha}{\tilde{\alpha}}\right)^{\zeta/2-1}\right] \|\hat{\nabla}_n f(x_n)\|_2$$  \hspace{1cm} (143)

**Proof.** Consider

$$\|h_n(\alpha) - x_n\|_2 \leq \|h_n(\alpha) - \Pi_x \left(x_n - \alpha H_n^{-1}\hat{\nabla}_n f(x_n), H_n\right)\|_2 + \|\Pi_x \left(x_n - \alpha H_n^{-1}\hat{\nabla}_n f(x_n), H_n\right) - x_n\|_2$$

It follows from (114) and (130) that

$$\|h_n(\alpha) - \Pi_x \left(x_n - \alpha H_n^{-1}\hat{\nabla}_n f(x_n), H_n\right)\|_2 \leq \frac{1}{\sqrt{\lambda_1}} \left\|h_n(\alpha) - \Pi_x \left(x_n - \alpha H_n^{-1}\hat{\nabla}_n f(x_n), H_n\right)\right\|_{H_n}$$

$$\leq \frac{3\alpha}{\sqrt{\lambda_1(1 - \eta_1)}} \left(\frac{\alpha}{\tilde{\alpha}}\right)^{\zeta/2-1} \left\|H_n^{-1}\hat{\nabla}_n f(x_n)\right\|_{H_n}$$

$$= \frac{3\alpha}{\sqrt{\lambda_1(1 - \eta_1)}} \left(\frac{\alpha}{\tilde{\alpha}}\right)^{\zeta/2-1} \sqrt{\hat{\nabla}_n f(x_n)^T H_n^{-1}\hat{\nabla}_n f(x_n)}$$

$$\leq \frac{3\alpha}{\lambda_1 \sqrt{1 - \eta_1}} \left(\frac{\alpha}{\tilde{\alpha}}\right)^{\zeta/2-1} \|\hat{\nabla}_n f(x_n)\|_2$$

Also, it was shown in Lemma 7.11 that

$$\left\|\Pi_x \left(x_n - \alpha H_n^{-1}\hat{\nabla}_n f(x_n), H_n\right) - x_n\right\|_2 \leq \frac{\alpha}{\lambda_1} \|\hat{\nabla}_n f(x_n)\|_2$$

\[ \square \]

**Lemma 7.12.** Consider a sequence $\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X}$ such that $x_n \to x \in \mathcal{X}$ as $n \to \infty$. Consider a corresponding sequence $\{h_n\}_{n \in \mathbb{N}}$ of arcs that satisfy (142), with $\{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P}$. Suppose that $f$ is continuously differentiable on a neighborhood of $\mathcal{X}$, and that $\{\hat{\nabla}_n f\}_{n \in \mathbb{N}}$ satisfies Assumption A 5.2. Then, for any sequence $\{\alpha_n\}_{n \in \mathbb{N}} \subset [0, \tilde{\alpha}]$, the sequence $\{h_n(\alpha_n)\}_{n \in \mathbb{N}}$ is bounded.
Proof. It follows from Lemma 7.11 that

\[ \|h_n(\alpha_n) - x_n\|_2 \leq \frac{\alpha_n}{\lambda_i} \left[ 1 + \frac{3}{\sqrt{1 - \eta_i}} \left( \frac{\alpha_n}{\alpha'} \right)^{\zeta/2 - 1} \right] \|\hat{\nabla}_n f(x_n)\|_2 \]

Since \( \hat{\nabla}_n f(x_n) \to \nabla f(x) \), \( \{\hat{\nabla}_n f(x_n)\}_{n \in \mathbb{N}} \) is bounded. Also, \( \{\alpha_n\}_{n \in \mathbb{N}} \) is bounded. Thus \( \{\|h_n(\alpha_n) - x_n\|_2\}_{n \in \mathbb{N}} \) is bounded. But \( x_n \to x \), and therefore \( \{h_n(\alpha_n)\}_{n \in \mathbb{N}} \) is bounded. \( \square \)

Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that \( x_n \to x \in \mathcal{X} \) as \( n \to \infty \). As before, let \( \mathcal{H} \) denote the set of limit points of the corresponding sequence \( \{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P} \). Recall that \( \mathcal{H} \) is a nonempty compact subset of \( \mathcal{P} \). As in Section 7.1.1, we wish to characterize the set \( \mathcal{Y}(\alpha) \) of limit points of the sequence \( \{h_n(\alpha)\}_{n \in \mathbb{N}} \). For any \( \alpha \in [0, \infty) \) and \( H \in \mathcal{H} \), define the set \( \mathcal{Y}(\alpha, H) \) similar to \( \mathcal{Y}(\alpha) \), as follows.

\[ \mathcal{Y}(\alpha, H) := \bigcup_{\{n_k\}_{k \in \mathbb{N}} \in \mathbb{N} : H_{n_k} \to H} \{y \in \mathcal{X} : y \text{ is a limit point of } \{h_{n_k}(\alpha)\}_{k \in \mathbb{N}}\} \quad (144) \]

Next we characterize the sets \( \mathcal{Y}(\alpha, H) \) for \( H \in \mathcal{H} \) and show that \( \mathcal{Y}(\alpha) = \bigcup_{H \in \mathcal{H}} \mathcal{Y}(\alpha, H) \).

Lemma 7.13. Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that \( x_n \to x \in \mathcal{X} \) as \( n \to \infty \). Consider a corresponding sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs that satisfy (142), with \( \{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P} \). Suppose that \( f \) is continuously differentiable on a neighborhood of \( \mathcal{X} \), and that \( \{\hat{\nabla}_n f\}_{n \in \mathbb{N}} \) satisfies Assumption A 5.2. Then, the following hold:

(a) For any \( \alpha \in [0, \alpha'] \) and \( H \in \mathcal{H} \),

\[ \emptyset \neq \mathcal{Y}(\alpha, H) \subset \hat{\Pi}_x(x, \alpha, -H^{-1}\nabla f(x), H) \]

(b) For any \( \alpha \in [0, \alpha'] \),

\[ \mathcal{Y}(\alpha) = \bigcup_{H \in \mathcal{H}} \mathcal{Y}(\alpha, H) \]

Proof. (a) Consider any \( \alpha \in [0, \alpha'] \) and \( H \in \mathcal{H} \). Let \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) be such that \( H_{n_k} \to H \) as \( k \to \infty \). It follows from Lemma 7.12 that the sequence \( \{h_{n_k}(\alpha)\}_{k \in \mathbb{N}} \) is bounded, and thus \( \mathcal{Y}(\alpha, H) \neq \emptyset \).

Next, we show that \( \mathcal{Y}(\alpha, H) \subset \hat{\Pi}_x(x, \alpha, -H^{-1}\nabla f(x), H) \). Consider any \( y \in \mathcal{Y}(\alpha, H) \), that is, \( y \) is a limit point of \( \{h_{n_k}(\alpha)\}_{k \in \mathbb{N}} \). Thus there is a further subsequence \( \{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}} \) such
that $h_{m_k}(\alpha) \to y$ as $k \to \infty$. Recall that $h_{m_k}(\alpha) \in \hat{\Pi}_x(x_{m_k}, \alpha, -H^{-1}_{m_k} \nabla f(x_{m_k}), H_{m_k})$, and thus it follows from (129) and (130) that for each $k \in \mathbb{N},$
\[
\frac{3\alpha}{\sqrt{1-\eta_1}} \left\| h_{m_k}(\alpha) - \Pi_x \left( x_{m_k} - \alpha H^{-1}_{m_k} \nabla f(x_{m_k}), H_{m_k} \right) \right\|_{H_{m_k}} \leq \left( \frac{\alpha}{\alpha'} \right)^{\zeta/2-1} \left\| H^{-1}_{m_k} \nabla f(x_{m_k}) \right\|_{H_{m_k}}
\]
and
\[
\left\| h_{m_k}(\alpha) - \Pi_x \left( x_{m_k} - \alpha H^{-1}_{m_k} \nabla f(x_{m_k}), H_{m_k} \right) \right\|_{H_{m_k}} \leq \frac{3\alpha}{\sqrt{1-\eta_1}} \left( \frac{\alpha}{\alpha'} \right)^{\zeta/2-1} \left\| H^{-1}_{m_k} \nabla f(x_{m_k}) \right\|_{H_{m_k}}
\]
Taking limits on both sides of the two inequalities above as $k \to \infty$, it follows from the continuity of $\Pi_x$ that
\[
\frac{3\alpha}{\sqrt{1-\eta_1}} \left\| y - \Pi_x \left( x - \alpha H^{-1} \nabla f(x), H \right) \right\|_H \left\| H^{-1} \nabla f(x) \right\|_H \leq \left( \frac{\alpha}{\alpha'} \right)^{\zeta/2-1} \left\| y - x \right\|_H
\]
and
\[
\left\| y - \Pi_x \left( x - \alpha H^{-1} \nabla f(x), H \right) \right\|_H \leq \frac{3\alpha}{\sqrt{1-\eta_1}} \left( \frac{\alpha}{\alpha'} \right)^{\zeta/2-1} \left\| H^{-1} \nabla f(x) \right\|_H
\]
Also, since $h_{m_k}(\alpha) \in \mathcal{X}$ for all $k \in \mathbb{N}$ and $\mathcal{X}$ is closed, it holds that $y \in \mathcal{X}$. Note that the two inequalities above show that (129) and (130) hold, and thus $y \in \hat{\Pi}_x(x, \alpha, -H^{-1} \nabla f(x), H).$
Therefore $\mathcal{Y}(\alpha, H) \subset \hat{\Pi}_x(x, \alpha, -H^{-1} \nabla f(x), H).$

(b) It follows from the definition of $\mathcal{Y}(\alpha, H)$ that
\[
\bigcup_{H \in \mathcal{H}} \mathcal{Y}(\alpha, H) \subset \mathcal{Y}(\alpha)
\]
Next, consider any $y \in \mathcal{Y}(\alpha)$. Then, there is a subsequence $\{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N}$ such that $h_{n_k}(\alpha) \to y$ as $k \to \infty$. Since $\{H_{n_k}\}_{k \in \mathbb{N}} \subset \mathcal{P}$, there is a further subsequence $\{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}}$, such that $H_{m_k} \to H \in \mathcal{H}$ as $k \to \infty$. Thus, $y \in \mathcal{Y}(\alpha, H)$ where $H \in \mathcal{H}$. Therefore,
\[
\mathcal{Y}(\alpha) \subset \bigcup_{H \in \mathcal{H}} \mathcal{Y}(\alpha, H)
\]
Since $\mathcal{H} \neq \emptyset$ and $\mathcal{Y}(\alpha, H) \neq \emptyset$ for every $H \in \mathcal{H}$, it follows that $\mathcal{Y}(\alpha)$ is nonempty.

Next, we show that $\mathcal{Y}(\alpha)$ is compact. Since $\mathcal{Y}(\alpha)$ is the set of limit points of $\{h_n(\alpha)\}_{n \in \mathbb{N}},$ $\mathcal{Y}(\alpha)$ is closed. Next we show that $\mathcal{Y}(\alpha)$ is bounded. Part a established that $\mathcal{Y}(\alpha, H) \subset \hat{\Pi}_x(x, \alpha, -H^{-1} \nabla f(x_n), H)$ for all $H \in \mathcal{H}$. Thus,
\[
\mathcal{Y}(\alpha) \subset \bigcup_{H \in \mathcal{H}} \hat{\Pi}_x(x, \alpha, -H^{-1} \nabla f(x_n), H)
\]
For any $H \in \mathcal{H}$ and $y \in \hat{\Pi}_x(x, \alpha, -H^{-1}\nabla f(x_n), H)$,

$$
\|y - x\|_2 \leq \|y - \Pi_x(x - \alpha H^{-1}\nabla f(x), H)\|_2 + \|\Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x\|_2
$$

It follows from (130) and (114) that

$$
\|y - \Pi_x(x - \alpha H^{-1}\nabla f(x), H)\|_2 \leq \frac{1}{\sqrt{\lambda_1}} \|y - \Pi_x(x - \alpha H^{-1}\nabla f(x), H)\|_H \leq \frac{3\alpha}{\sqrt{\lambda_1(1 - \eta_1)}} \left(\frac{\alpha}{\alpha'}\right)^{\zeta/2 - 1} \|H^{-1}\nabla f(x)\|_H
$$

(146)

Also, it follows from (114) and (117) in Property P 7.4 that

$$
\|\Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x\|_2 \leq \frac{1}{\sqrt{\lambda_1}} \|\Pi_x(x - \alpha H^{-1}\nabla f(x), H) - x\|_H \leq \frac{3\alpha}{\sqrt{\lambda_1(1 - \eta_1)}} \left(\frac{\alpha}{\alpha'}\right)^{\zeta/2 - 1} \|H^{-1}\nabla f(x)\|_H
$$

(147)

Combining (146) and (147), it follows that

$$
\|y - x\|_2 \leq \left[\frac{3}{\sqrt{1 - \eta_1}} \left(\frac{\alpha}{\alpha'}\right)^{\zeta/2 - 1} + 1\right] \frac{\alpha}{\sqrt{\lambda_1}} \|H^{-1}\nabla f(x)\|_H
$$

Since $H \in \mathcal{H} \subset \mathcal{P}$, it follows from (114) that

$$
\|H^{-1}\nabla f(x)\|_H = \sqrt{\nabla f(x)^T H^{-1}\nabla f(x)} \leq \frac{1}{\sqrt{\lambda_1}} \|\nabla f(x)\|_2
$$

Thus, for any $H \in \mathcal{H}$ and $y \in \hat{\Pi}_x(x, \alpha, -H^{-1}\nabla f(x_n), H)$,

$$
\|y - x\|_2 \leq \left[\frac{3}{\sqrt{1 - \eta_1}} \left(\frac{\alpha}{\alpha'}\right)^{\zeta/2 - 1} + 1\right] \frac{\alpha}{\sqrt{\lambda_1}} \|\nabla f(x)\|_2
$$

(148)

The right side of (148) is independent of $H$ and $y$, and thus

$$
\sup_{H \in \mathcal{H}} \sup_{y \in \hat{\Pi}_x(x, \alpha, -H^{-1}\nabla f(x_n), H)} \|y - x\|_2 \leq \left[\frac{3}{\sqrt{1 - \eta_1}} \left(\frac{\alpha}{\alpha'}\right)^{\zeta/2 - 1} + 1\right] \frac{\alpha}{\lambda_1} \|\nabla f(x)\|_2
$$

(149)

Hence, $\cup_{H \in \mathcal{H}} \hat{\Pi}_x(x, \alpha, -H^{-1}\nabla f(x_n), H)$ is bounded, and therefore $\mathcal{Y}(\alpha)$ is bounded.

\[\Box\]

**Lemma 7.14.** Consider a sequence $\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X}$ such that $x_n \to x \in \mathcal{X}$ as $n \to \infty$. Consider a corresponding sequence $\{h_n\}_{n \in \mathbb{N}}$ of arcs that satisfy (142), with $\{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P}$. Suppose that $f$ is continuously differentiable on a neighborhood of $\mathcal{X}$, and that $\hat{\nabla}_n f$ satisfies Assumption A 5.2. Let $\mathcal{S}$ denote the set of stationary points of $f$ on $\mathcal{X}$, as defined in (109). Then, the following properties hold:
P 7.13. For any $\alpha \in [0, \bar{\alpha}')$, the sequence $\{h_n(\alpha)\}_{n \in \mathbb{N}}$ is bounded, and $\mathcal{Y}(\alpha)$ is nonempty and compact.

P 7.14. For any sequence $\{\alpha_n\}_{n \in \mathbb{N}} \subset [0, \infty)$ such that $\alpha_n \to 0$ as $n \to \infty$, it holds that

$$\lim_{n \to \infty} h_n(\alpha_n) = x$$

P 7.15.

$$\lim_{\alpha \to 0} \sup_{y \in \mathcal{Y}(\alpha)} \|y - x\|_2 = 0$$

P 7.16. If $x \notin \mathcal{S}$, then for any $\alpha \in (0, \bar{\alpha}')$ and any $y \in \mathcal{Y}(\alpha)$,

$$\|y - x\|_2 > 0$$

In addition, if $\bar{\alpha}$ satisfies

$$0 < \bar{\alpha} \leq \left(\frac{1 - \eta_1}{4}\right)^{\frac{1}{\alpha}} \bar{\alpha}'$$

then

$$\nabla f(x)^T (y - x) \leq -\frac{\eta_1 \lambda_1}{\alpha} \| y - x \|_2^2 < 0$$

for all $\alpha \in (0, \bar{\alpha}]$. Also,

$$\sup_{\alpha \in (0, \bar{\alpha}]} \sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^T \frac{y - x}{\| y - x \|_2} < 0$$

P 7.17. If $x \in \mathcal{S}$, then $\mathcal{Y}(\alpha) = \{x\}$ for every $\alpha \in [0, \bar{\alpha}')$. Consequently, $\nabla f(x)^T (y - x) = 0$ for all $\alpha \in [0, \bar{\alpha}')$ and $y \in \mathcal{Y}(\alpha)$.

Proof. (a) Property P 7.13: It follows from Lemma 7.12 that for any $\alpha \in [0, \bar{\alpha}')$, the sequence $\{h_n(\alpha)\}_{n \in \mathbb{N}}$ is bounded. Hence $\mathcal{Y}(\alpha)$ is nonempty and compact.

(b) Property P 7.14: It follows from Lemma 7.11 that

$$\|h_n(\alpha_n) - x_n\|_2 \leq \frac{\alpha_n}{\lambda_1} \left[ 1 + \frac{3}{\sqrt{1 - \eta_1}} \left(\frac{\alpha_n}{\bar{\alpha}'}\right)^{\zeta/2 - 1} \right] \|\hat{\nabla}_n f(x_n)\|_2$$

Since $\hat{\nabla}_n f(x_n) \to \nabla f(x)$ and $\alpha_n \to 0$, it follows that

$$\|h_n(\alpha_n) - x_n\|_2 \to 0$$

as $n \to \infty$. Since $x_n \to x$, it follows that $h_n(\alpha_n) \to x$ as $n \to \infty$.

(c) Property P 7.15: Property P 7.15 follows from Property P 7.13, Property P 7.14, and Lemma 7.1.
(d) Property P 7.16: Consider any $\alpha \in (0, \bar{\alpha}']$ and any $y \in \mathcal{Y}(\alpha)$. It follows from Part b of Lemma 7.13 that $y \in \mathcal{Y}(\alpha, H)$ for some $H \in \mathcal{H} \subset S_{++}^{l \times l}$. Part a of Lemma 7.13 established that $\mathcal{Y}(\alpha, H) \subset \Pi_{x}(x, \alpha, -H^{-1}\nabla f(x), H)$ for every $\alpha \in [0, \bar{\alpha}]$ and $H \in \mathcal{H}$. Since $x \notin \mathcal{S}$, it follows from (137) in Lemma 7.10 and (114) that $\|y - x\|_{2} \geq \|y - x\|_{H} / \sqrt{\lambda_{2}} > 0$.

Next, suppose that $0 < \bar{\alpha} \leq [(1 - \eta_{1})/4]^{1/(c-2)} \bar{\alpha}'$, and consider any $\alpha \in (0, \bar{\alpha}]$ and any $y \in \mathcal{Y}(\alpha)$. As before, $y \in \mathcal{Y}(\alpha, H) \subset \Pi_{x}(x, \alpha, -H^{-1}\nabla f(x), H)$ for some $H \in \mathcal{H} \subset \mathcal{P} \subset S_{++}^{l \times l}$. It follows from Corollary 7.8 and (114) that

$$\nabla f(x)^{T}(y - x) \leq -\frac{\eta_{1}}{\alpha} \|y - x\|_{2}^{2} \leq -\frac{\eta_{1}}{\alpha} \sqrt{\lambda_{1}} \|y - x\|_{2} \|y - x\|_{H} \leq -\frac{\eta_{1}}{\alpha} \lambda_{1} \|y - x\|_{2}^{2}$$

Since $\|y - x\|_{2} > 0$, it follows that

$$\nabla f(x)^{T} \frac{y - x}{\|y - x\|_{2}} \leq -\frac{\eta_{1}}{\alpha} \sqrt{\lambda_{1}} \frac{\Pi_{x}(x - \alpha H^{-1}\nabla f(x), H) - x}{H}$$

It follows from (136) and (119) of Property P 7.4 that

$$\frac{y - x}{\|y - x\|_{2}} \leq -\frac{\eta_{1}}{\alpha} \sqrt{\lambda_{1}} \frac{\Pi_{x}(x - \bar{\alpha} H^{-1}\nabla f(x), H) - x}{\bar{\alpha}}$$

Thus,

$$\sup_{y \in \mathcal{Y}(\alpha, H)} \nabla f(x)^{T} \frac{y - x}{\|y - x\|_{2}} \leq -\frac{\eta_{1}}{\alpha} \sqrt{\lambda_{1}} \frac{\inf_{H \in \mathcal{H}} \|\Pi_{x}(x - \bar{\alpha} H^{-1}\nabla f(x), H) - x\|_{H}}{2\bar{\alpha}}$$

Since $\mathcal{Y}(\alpha) = \bigcup_{H \in \mathcal{H}} \mathcal{Y}(\alpha, H)$, it follows that

$$\sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^{T} \frac{y - x}{\|y - x\|_{2}} = \sup_{H \in \mathcal{H}} \sup_{y \in \mathcal{Y}(\alpha, H)} \nabla f(x)^{T} \frac{y - x}{\|y - x\|_{2}} \leq -\frac{\eta_{1}}{\alpha} \sqrt{\lambda_{1}} \inf_{H \in \mathcal{H}} \|\Pi_{x}(x - \bar{\alpha} H^{-1}\nabla f(x), H) - x\|_{H}$$

Since $\mathcal{H} \subset S_{++}^{l \times l}$ and $x \notin \mathcal{S}$, it follows from Property P 7.6 that $\|\Pi_{x}(x - \bar{\alpha} H^{-1}\nabla f(x), H) - x\|_{H} > 0$ for all $H \in \mathcal{H}$. But $\mathcal{H}$ is nonempty and compact, and $\|\Pi_{x}(x - \bar{\alpha} H^{-1}\nabla f(x), H) - x\|_{H}$ is a continuous function of $H$. Hence,

$$\inf_{H \in \mathcal{H}} \|\Pi_{x}(x - \bar{\alpha} H^{-1}\nabla f(x), H) - x\|_{H} > 0$$

Therefore,

$$\sup_{\alpha \in (0, \bar{\alpha}]} \sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^{T} \frac{y - x}{\|y - x\|_{2}} < 0$$

90
(e) Property P 7.17: Part b of Lemma 7.13 established that \( \mathcal{Y}(\alpha) = \bigcup_{H \in \mathcal{H}} \mathcal{Y}(\alpha, H) \) for each \( \alpha \in [0, \alpha'] \), and thus it is sufficient to show that \( \mathcal{Y}(\alpha, H) = \{x\} \) for all \( \alpha \in [0, \alpha'] \) and \( H \in \mathcal{H} \subset \mathbb{S}_{++}^{l \times l} \).

Also, Part a of Lemma 7.13 established that \( \mathcal{Y}(\alpha, H) \subset \hat{\Pi}_x(x, \alpha, -H^{-1}\nabla f(x), H) \) for every \( \alpha \in [0, \alpha'] \) and \( H \in \mathcal{H} \). Part b of Lemma 7.10 showed that if \( x \in \mathcal{S} \), then \( \hat{\Pi}_x(x, \alpha, -H^{-1}\nabla f(x), H) = \{x\} \) for all \( \alpha \in [0, \alpha'] \) and \( H \in \mathbb{S}_{++}^{l \times l} \). Therefore, \( \mathcal{Y}(\alpha, H) = \{x\} \) for all \( \alpha \in [0, \alpha'] \) and \( H \in \mathcal{H} \).

\[ \square \]

Lemma 7.15. Consider a sequence \( \{\alpha_n\}_{n \in \mathbb{N}} \subset (0, \alpha'] \) such that \( \alpha_n \to 0 \) as \( n \to \infty \), and a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that \( x_n \to x \in \mathcal{X} \) as \( n \to \infty \). Consider a corresponding sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs that satisfy (142), with \( \{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P} \). Suppose that \( f \) is continuously differentiable on a neighborhood of \( \mathcal{X} \), and that \( \{\nabla f\}_{n \in \mathbb{N}} \) satisfies Assumption A 5.2. Let \( \mathcal{S} \) denote the set of stationary points of \( f \) on \( \mathcal{X} \), as defined in (109). If

\[ \lim_{n \to \infty} \frac{\|h_n(\alpha_n) - x_n\|_2}{\alpha_n} = 0 \]

then \( x \in \mathcal{S} \).

Proof. Consider

\[ \left\| \Pi_x(x_n - \alpha_n H_n^{-1} \hat{\nabla}_n f(x_n), H_n) - x_n \right\|_2 \leq \left\| \Pi_x(x_n - \alpha_n H_n^{-1} \hat{\nabla}_n f(x_n), H_n) - h_n(\alpha_n) \right\|_2 + \|h_n(\alpha_n) - x_n\|_2 \quad (150) \]

It follows from (130) and (114) that

\[ \left\| \Pi_x(x_n - \alpha_n H_n^{-1} \hat{\nabla}_n f(x_n), H_n) - h_n(\alpha_n) \right\|_2 \leq \frac{1}{\sqrt{\lambda_1}} \left\| \Pi_x(x_n - \alpha_n H_n^{-1} \hat{\nabla}_n f(x_n), H_n) - h_n(\alpha_n) \right\|_{H_n} \]

\[ \leq \frac{3}{\sqrt{\lambda_1(1 - \eta)}} \left( \frac{\alpha_n}{\alpha'} \right)^{\gamma/2 - 1} \left\| H_n^{-1} \hat{\nabla}_n f(x_n) \right\|_{H_n} \]

\[ = \frac{3}{\sqrt{\lambda_1(1 - \eta)}} \left( \frac{\alpha_n}{\alpha'} \right)^{\gamma/2 - 1} \left\| \hat{\nabla}_n f(x_n)^T H_n^{-1} \hat{\nabla}_n f(x_n) \right\|_2 \]

\[ \leq \frac{3}{\lambda_1(1 - \eta)} \left( \frac{\alpha_n}{\alpha'} \right)^{\gamma/2 - 1} \left\| \hat{\nabla}_n f(x_n) \right\|_2 \]

Recall that \( \hat{\nabla}_n f(x_n) \to \nabla f(x) \) and \( \alpha_n \to 0 \). Thus

\[ \lim_{n \to \infty} \left\| \Pi_x(x_n - \alpha_n H_n^{-1} \hat{\nabla}_n f(x_n), H_n) - h_n(\alpha_n) \right\|_2 = 0 \]
Hence it follows from (150) that
\[
\lim_{n \to \infty} \frac{\| \Pi_X(x_n) - \alpha_n H_n^{-1} \nabla_n f(x_n), H_n \| - x_n \|_2}{\alpha_n} = 0
\]
Then it follows from Lemma 7.6 that \( x \in S \).

Next, we illustrate a practical method to find \( y \in \hat{\Pi}_X(x, \alpha, d, H) \) for given \( x \in X \), \( \alpha \in [0, \bar{\alpha}] \), \( d \in \mathbb{R}^l \), and \( H \in \mathbb{S}_+^{l \times l} \), that is, to find \( y \in X \) that satisfies the inequalities (129) and (130). Note that the term \( \| y - \Pi_X(x + \alpha d, H) \|_H \) occurs on the left side of both inequalities. Thus, in order to check whether these two inequalities are satisfied by a candidate point \( y \in X \), we need to estimate the value of \( \| y - \Pi_X(x + \alpha d, H) \|_H \). This can be done as follows.

Recall that \( \Pi_X(x + \alpha d, H) \) is the unique optimal solution of the problem
\[
\min_{y \in X} \| x + \alpha d - y \|_H^2
\]
Suppose that we use an algorithm \( \mathbb{P} \) that generates a sequence \( \{ y^k \}_{k \in \mathbb{N}} \subset X \) of feasible approximations of \( \Pi_X(x + \alpha d, H) \). We use superscripts to denote the elements of this sequence to distinguish them from the sequence \( \{ x_n \}_{n \in \mathbb{N}} \) of iterates generated by a line search algorithm. Suppose that \( \mathbb{P} \) also generates a duality gap \( \delta^k \) at each iteration \( k \). Then we can use (129) and (130) together as a stopping criterion in \( \mathbb{P} \). From the definition of a duality gap,
\[
\| x + \alpha d - y^k \|_H^2 - \delta^k \leq \| x + \alpha d - \Pi_X(x + \alpha d, H) \|_H^2
\]
for each iteration \( k \). That is,
\[
2[ x + \alpha d - \Pi_X(x + \alpha d, H)]^T H[ \Pi_X(x + \alpha d, H) - y^k] + \| \Pi_X(x + \alpha d, H) - y^k \|_H^2 \leq \delta^k
\]
Since \( y^k \in X \), it follows from Property P 7.2 that \( 2[ x + \alpha d - \Pi_X(x + \alpha d, H)]^T H[ \Pi_X(x + \alpha d, H) - y^k] \geq 0 \). Thus,
\[
\| \Pi_X(x + \alpha d, H) - y^k \|_H \leq \sqrt{\delta^k}
\]
Hence, algorithm \( \mathbb{P} \) can stop at an iteration \( j \) if
\[
\frac{3\alpha}{\sqrt{1 - \eta_1}} \sqrt{\delta^j} \| d \|_H \leq \left( \frac{\alpha}{\alpha^j} \right)^{\zeta/2 - 1} \| y^j - x \|_H^2 \quad \text{and} \quad \sqrt{\delta^j} \leq \frac{3\alpha}{\sqrt{1 - \eta_1}} \left( \frac{\alpha}{\alpha^j} \right)^{\zeta/2 - 1} \| d \|_H
\]
7.1.3 Feasible Direction Methods

When the arc rules based on the projection and inexact projection discussed in the previous sections are used, a projection problem

\[
\min_{y \in \mathcal{X}} \left( x_n - \alpha H_n^{-1} \nabla f(x_n) - y \right)^T H_n \left( x_n - \alpha H_n^{-1} \nabla f(x_n) - y \right)
\]  

(153)

has to be solved (at least approximately) at a number of step sizes \( \alpha \) at each iteration \( n \). Even if the set \( \mathcal{X} \) is polyhedral, the projection problem (153) is a quadratic program which can be computationally expensive enough to make the method less efficient than methods based on other arc rules. In this section, we describe arc rules for which the entire arc \( \{ h_n(\alpha) : \alpha \in [0, \bar{\alpha}] \} \) is known in closed form after solving at most one projection problem as in (153).

At an iteration \( n \), given current iterate \( x_n \in \mathcal{X} \), we first choose a target point \( \Phi_n \in \mathcal{X} \). The method for choosing a target point is called a target point rule. Then, the arc \( h_n : [0, 1] \mapsto \mathcal{X} \) is defined as follows:

\[
h_n(\alpha) := x_n + \alpha (\Phi_n - x_n)
\]  

(154)

That is, the arc \( \{ h_n(\alpha) : \alpha \in [0, 1] \} \) is the line segment \([x_n, \Phi_n]\). Since \( \mathcal{X} \) is convex, \( h_n(\alpha) \in \mathcal{X} \) for all \( x_n \in \mathcal{X} \), \( \Phi_n \in \mathcal{X} \), and \( \alpha \in [0, 1] \). In other words, \( \Phi_n - x_n \) is a feasible direction for \( \mathcal{X} \) at the point \( x_n \). Accordingly, we refer to line search methods that use this class of arc rules as feasible direction methods. Also, note that \( h_n(x_n, 0) = x_n \).

Next we consider some specific examples of target point rules.

**Example 7.1.** Suppose that the set \( \mathcal{X} \) is compact. At an iteration \( n \), given current iterate \( x_n \in \mathcal{X} \), the Franke-Wolfe or conditional gradient method chooses \( \Phi_n \) as follows: If \( \nabla f(x_n) \) can be evaluated easily, then

\[
\Phi_n \in \arg\min_{y \in \mathcal{X}} \nabla f(x_n)^T (y - x_n)
\]  

(155)

If \( \nabla f(x_n) \) cannot be evaluated easily, then we use \( \hat{\nabla}_n f(x_n) \) and choose \( \Phi_n \) as follows:

\[
\Phi_n \in \arg\min_{y \in \mathcal{X}} \hat{\nabla}_n f(x_n)^T (y - x_n)
\]  

(156)

The set \( \arg\min \{ d^T (y - x) : y \in \mathcal{X} \} \) is nonempty for each \( x, d \in \mathbb{R}^l \) since \( \mathcal{X} \) is compact. This method is often used when the set \( \mathcal{X} \) is polyhedral since \( \Phi_n \) can be easily obtained by solving a linear program.
Example 7.2. At an iteration \( n \), given current iterate \( x_n \in X \), choose a constant \( \gamma > 0 \), and a matrix \( H_n \in \mathcal{P} \). If \( \nabla f(x_n) \) can be evaluated exactly, then choose \( \Phi_n \) as follows:

\[
\Phi_n := \Pi_x(x_n - \gamma H_n^{-1} \nabla f(x_n), H_n) \tag{157}
\]

If \( \nabla f(x_n) \) cannot be evaluated exactly, then we use \( \hat{\nabla} f(x_n) \) and choose \( \Phi_n \) as follows:

\[
\Phi_n := \Pi_x(x_n - \gamma H_n^{-1} \hat{\nabla} f(x_n), H_n) \tag{158}
\]

Example 7.3. Instead of using exact projections as in Example 7.2, we can use inexact projections to choose target points. Choose the constants \( \bar{\alpha}' > 0 \), \( \eta_1 \in (0,1) \) and \( \zeta > 2 \) and

\[
0 < \gamma \leq \left( \frac{1 - \eta_1}{4} \right)^{\frac{1}{\zeta - 2}} \bar{\alpha}'
\]

At an iteration \( n \), given current iterate \( x_n \in X \), choose a matrix \( H_n \in \mathcal{P} \). If \( \nabla f(x_n) \) can be evaluated exactly, then choose \( \Phi_n \) as follows:

\[
\Phi_n \in \Pi_x(x_n, \gamma, -H_n^{-1} \nabla f(x_n), H_n) \tag{159}
\]

If \( \nabla f(x_n) \) cannot be evaluated exactly, then we use \( \hat{\nabla} f(x_n) \) and choose \( \Phi_n \) as follows:

\[
\Phi_n \in \Pi_x(x_n, \gamma, -H_n^{-1} \hat{\nabla} f(x_n), H_n) \tag{160}
\]

Next we consider some properties of the arc rules defined in (154).

Lemma 7.16. Consider a sequence \( \{x_n\}_{n\in\mathbb{N}} \subset X \) such that \( x_n \to x \in X \) as \( n \to \infty \). Consider a corresponding sequence \( \{\Phi_n\}_{n\in\mathbb{N}} \) of target points and sequence \( \{h_n\}_{n\in\mathbb{N}} \) of arcs defined in (154). Assume that the sequence \( \{\Phi_n\}_{n\in\mathbb{N}} \) is bounded. Then, for any sequence \( \{\alpha_n\}_{n\in\mathbb{N}} \subset [0,1] \), the sequence \( \{h_n(\alpha_n)\}_{n\in\mathbb{N}} \) is bounded.

Proof. Since \( x_n \to x \), \( \{x_n\}_{n\in\mathbb{N}} \) is bounded. Also, \( \{\alpha_n\}_{n\in\mathbb{N}} \) and \( \{\Phi_n\}_{n\in\mathbb{N}} \) are bounded. Thus \( \{h_n(\alpha_n)\}_{n\in\mathbb{N}} \), with \( h_n(\alpha_n) := x_n + \alpha_n(\Phi_n - x_n) \) is bounded. \( \square \)

Consider any sequence \( \{x_n\}_{n\in\mathbb{N}} \subset X \). As before, for any \( \alpha \geq 0 \), let \( \mathcal{Y}(\alpha) \) denote the set of limit points of the sequence \( \{h_n(\alpha)\}_{n\in\mathbb{N}} \). Let \( \mathcal{Y}_\Phi \) denote the set of limit points of the sequence \( \{\Phi_n\}_{n\in\mathbb{N}} \). Lemmas 7.17 and 7.18 characterize \( \mathcal{Y}(\alpha) \) in terms of \( \mathcal{Y}_\Phi \). Subsequently, we characterize \( \mathcal{Y}_\Phi \) for each of the specific examples of target points discussed earlier.
Lemma 7.17. Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset X \) such that \( x_n \to x \in X \) as \( n \to \infty \). Consider a corresponding sequence \( \{\Phi_n\}_{n \in \mathbb{N}} \) of target points and sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs defined in (154). Assume that the sequence \( \{\Phi_n\}_{n \in \mathbb{N}} \) is bounded. Then, for any \( \alpha \in [0, 1] \), the set \( \mathcal{Y}(\alpha) \) of limit points of the sequence \( \{h_n(\alpha)\}_{n \in \mathbb{N}} \) satisfies

\[
\mathcal{Y}(\alpha) = \{ x + \alpha(z-x) : z \in \Phi \}
\]

Proof. Consider any \( z \in \Phi \). Then, there is a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( \Phi_{n_k} \to z \) as \( k \to \infty \). For any \( \alpha \in [0, 1] \),

\[
h_{n_k}(\alpha) = x_{n_k} + \alpha(\Phi_{n_k} - x_{n_k})
\]

Thus,

\[
\lim_{k \to \infty} h_{n_k}(x_{n_k}, \alpha) = x + \alpha(z-x)
\]

That is, \( x + \alpha(z-x) \in \mathcal{Y}(\alpha) \) for all \( z \in \Phi \) and all \( \alpha \in [0, 1] \).

Next, consider any \( \alpha \in [0, 1] \) and any \( y \in \mathcal{Y}(\alpha) \). Then, there is a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( h_{n_k}(\alpha) \to y \) as \( k \to \infty \). Since \( \{\Phi_{n_k}\}_{k \in \mathbb{N}} \) is bounded, there is a further subsequence \( \{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}} \) such that \( \Phi_{m_k} \to z \in \Phi \) as \( k \to \infty \). Then,

\[
y = \lim_{k \to \infty} h_{m_k}(\alpha) = \lim_{k \to \infty} [x_{m_k} + \alpha(\Phi_{m_k} - x_{m_k})] = x + \alpha(z-x)
\]

Therefore, \( \mathcal{Y}(\alpha) = \{ x + \alpha(z-x) : z \in \Phi \} \).

Lemma 7.18. Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset X \) such that \( x_n \to x \in X \) as \( n \to \infty \). Consider a corresponding sequence \( \{\Phi_n\}_{n \in \mathbb{N}} \) of target points and sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs defined in (154). Assume that the sequence \( \{\Phi_n\}_{n \in \mathbb{N}} \) is bounded. Suppose that \( f \) is continuously differentiable at \( x \). Let \( S \) denote the set of stationary points of \( f \) on \( X \), as defined in (109). Then, the following properties hold:

P 7.18. For any \( \alpha \in [0, 1] \), the sequence \( \{h_n(\alpha)\}_{n \in \mathbb{N}} \) is bounded, and \( \mathcal{Y}(\alpha) \) is nonempty and compact.

P 7.19. For any sequence \( \{\alpha_n\}_{n \in \mathbb{N}} \subset [0, 1] \) such that \( \alpha_n \to 0 \) as \( n \to \infty \), it holds that

\[
\lim_{n \to \infty} h_n(\alpha_n) = x
\]

P 7.20.

\[
\lim_{\alpha \downarrow 0} \sup_{y \in \mathcal{Y}(\alpha)} \|y - x\|_2 = 0
\]

95
P 7.21. If

\[ \sup_{z \in \mathcal{Y}_\Phi} \nabla f(x)^T (z - x) < 0 \]

then for any \( \alpha \in (0, 1] \) and any \( y \in \mathcal{Y}(\alpha) \),

\[ \| y - x \|_2 > 0 \]

and

\[ \sup_{\alpha \in (0, 1]} \sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^T \frac{y - x}{\| y - x \|_2} < 0 \]

P 7.22. If \( \nabla f(x)^T (z - x) \leq 0 \) for all \( z \in \mathcal{Y}_\Phi \), then

\[ \nabla f(x)^T (y - x) \leq 0 \]

for all \( \alpha \in [0, 1] \) and all \( y \in \mathcal{Y}(\alpha) \).

Proof. (a) **Property P 7.18:** It follows from Lemma 7.16 that for any \( \alpha \in [0, 1] \), the sequence \( \{ h_n(\alpha) \} \) is bounded. Hence \( \mathcal{Y}(\alpha) \) is nonempty and compact.

(b) **Property P 7.19:** Note that

\[ \| h_n(\alpha_n) - x_n \|_2 = \alpha_n \| \Phi_n - x_n \|_2 \]

Also, \( \{ \Phi_n \} \) and \( \{ x_n \} \) are both bounded sequences. Therefore,

\[ \lim_{n \to \infty} \| h_n(\alpha_n) - x_n \|_2 = \lim_{n \to \infty} \alpha_n \| \Phi_n - x_n \|_2 = 0 \]

Since \( x_n \to x \), it follows that \( h_n(\alpha_n) \to x \).

(c) **Property P 7.20:** Property P 7.20 follows from Property P 7.18, Property P 7.19, and Lemma 7.1.

(d) **Property P 7.21:** It follows from the assumption that \( \sup_{z \in \mathcal{Y}_\Phi} \nabla f(x)^T (z - x) < 0 \) that

\[ \inf_{z \in \mathcal{Y}_\Phi} \| z - x \|_2 > 0 \]

Then it follows from Lemma 7.17 that for any \( \alpha \in (0, 1] \),

\[ \inf_{y \in \mathcal{Y}(\alpha)} \| y - x \|_2 = \alpha \inf_{z \in \mathcal{Y}_\Phi} \| z - x \|_2 > 0 \]
Also, it follows from Lemma 7.17 that for any \( \alpha \in (0, 1] \) and any \( y \in \mathcal{Y}(\alpha) \), there is \( z \in \mathcal{Y}_\phi \) such that
\[
\nabla f(x)^T \frac{y - x}{\|y - x\|_2} = \nabla f(x)^T \frac{z - x}{\|z - x\|_2}
\]
and vice versa. Thus, for each \( \alpha \in (0, 1] \),
\[
\sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^T \frac{y - x}{\|y - x\|_2} = \sup_{z \in \mathcal{Y}_\phi} \nabla f(x)^T \frac{z - x}{\|z - x\|_2}
\]
Since the right side of the above equation does not depend on \( \alpha \), it follows that
\[
\sup_{\alpha \in (0, 1]} \sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^T \frac{y - x}{\|y - x\|_2} = \sup_{z \in \mathcal{Y}_\phi} \nabla f(x)^T \frac{z - x}{\|z - x\|_2}
\]
It follows from the assumption that
\[
\sup_{z \in \mathcal{Y}_\phi} \nabla f(x)^T (z - x) < 0
\]
and \( \mathcal{Y}_\phi \) being bounded that
\[
\sup_{z \in \mathcal{Y}_\phi} \nabla f(x)^T \frac{z - x}{\|z - x\|_2} < 0
\]
(e) **Property P 7.22:** It follows from Lemma 7.17 that for any \( \alpha \in [0, 1] \) and \( y \in \mathcal{Y}(\alpha) \), there is \( z \in \mathcal{Y}_\phi \) such that
\[
\nabla f(x)^T (y - x) = \alpha \nabla f(x)^T (z - x) \leq 0
\]
\[\square\]

**Lemma 7.19.** Suppose that the set \( \mathcal{X} \) is compact. Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that \( x_n \to x \in \mathcal{X} \) as \( n \to \infty \). Consider a corresponding sequence \( \{\Phi_n\}_{n \in \mathbb{N}} \) of target points defined in (156). Suppose that \( f \) is continuously differentiable on a neighborhood of \( \mathcal{X} \), and that \( \{\hat{\nabla}_n f\}_{n \in \mathbb{N}} \) satisfies Assumption A 5.2. Let \( S \) denote the set of stationary points of \( f \) on \( \mathcal{X} \), as defined in (109). Then

1. The set \( \mathcal{Y}_\phi \) of limit points of the sequence \( \{\Phi_n\}_{n \in \mathbb{N}} \) satisfies
\[
\emptyset \neq \mathcal{Y}_\phi \subset \arg \min_{y \in \mathcal{X}} \nabla f(x)^T (y - x)
\]
2. If \( x \notin \mathcal{S} \), then
\[
\sup_{z \in \mathcal{Y}_\phi} \nabla f(x)^T (z - x) < 0
\]

97
If \( x \in S \), then
\[
\sup_{z \in \mathcal{Y}_\Phi} \nabla f(x)^T (z - x) = 0
\]

Consequently, the sequence \( \{h_n\}_{n \in \mathbb{N}} \) of feasible direction arcs defined in (154) satisfy Properties P 7.21 and P 7.22.

Proof. 1. Compactness of \( \mathcal{X} \) implies that \( \mathcal{Y}_\Phi \) is nonempty and compact. Consider any \( z \in \mathcal{Y}_\Phi \). Then, there is a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subseteq \mathbb{N} \) such that \( \Phi_{n_k} \to z \) as \( k \to \infty \). It follows from the definition of \( \Phi_n \) in (156) that
\[
\hat{\nabla}_{n_k} f(x_{n_k})^T (\Phi_{n_k} - x_{n_k}) \leq \hat{\nabla}_{n_k} f(x_{n_k})^T (y - x_{n_k}) \quad \text{for all} \quad y \in \mathcal{X}
\]
Taking the limit as \( k \to \infty \), it follows that
\[
\nabla f(x)^T (z - x) \leq \nabla f(x)^T (y - x) \quad \text{for all} \quad y \in \mathcal{X}
\]
Therefore, \( z \in \arg \min_{y \in \mathcal{X}} \nabla f(x)^T (y - x) \).

2. It follows from the definition of \( S \) in (109) that if \( x \notin S \), then
\[
\min_{y \in \mathcal{X}} \nabla f(x)^T (y - x) < 0
\]
Since \( \mathcal{Y}_\Phi \subseteq \arg \min_{y \in \mathcal{X}} \nabla f(x)^T (y - x) \), it follows that
\[
\sup_{z \in \mathcal{Y}_\Phi} \nabla f(x)^T (z - x) < 0
\]
Also, if \( x \in S \), then
\[
\min_{y \in \mathcal{X}} \nabla f(x)^T (y - x) = 0
\]
and thus
\[
\sup_{z \in \mathcal{Y}_\Phi} \nabla f(x)^T (z - x) = 0
\]

Lemma 7.20. Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that \( x_n \to x \in \mathcal{X} \) as \( n \to \infty \). Consider a corresponding sequence \( \{\Phi_n\}_{n \in \mathbb{N}} \) of target points defined in (158) for some \( \gamma > 0 \) and \( \{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P} \). Suppose that \( f \) is continuously differentiable on a neighborhood of \( \mathcal{X} \), and that \( \{\hat{\nabla}_n\}_{n \in \mathbb{N}} \) satisfies Assumption A 5.2. Let \( S \) denote the set of stationary points of \( f \) on \( \mathcal{X} \), as defined in (109). Let \( \mathcal{H} \) denote the set of limit points of the sequence \( \{H_n\}_{n \in \mathbb{N}} \). Then the following properties hold.
1. The sequence \( \{ \Phi_n \}_{n \in \mathbb{N}} \) of target points is bounded. Hence the set \( \mathcal{Y}_\Phi \) of limit points of the sequence \( \{ \Phi_n \}_{n \in \mathbb{N}} \) is nonempty and compact.

2. The set \( \mathcal{Y}_\Phi \) satisfies
\[
\mathcal{Y}_\Phi = \{ \Pi_x(x - \gamma H^{-1} \nabla f(x), H) : H \in \mathcal{H} \}
\]

3. If \( x \notin S \), then
\[
\nabla f(x)^T (z - x) \leq \frac{\lambda_1}{\gamma} \| z - x \|^2_2 < 0
\]
\[
\sup_{z \in \mathcal{Y}_\Phi} \nabla f(x)^T (z - x) < 0
\]

4. If \( x \in S \), then
\[
\mathcal{Y}_\Phi = \{ x \}
\]

Consequently, the sequence \( \{ h_n \}_{n \in \mathbb{N}} \) of feasible direction arcs defined in (154) satisfy Properties P 7.21 and P 7.22. In addition, if \( x \in S \), then for all \( \alpha \in [0,1] \),
\[
\mathcal{Y}(\alpha) = \{ x \}
\]

Proof. 1. It follows from Lemma 7.3 that for any \( \gamma \in [0, \infty) \), the sequence \( \{ \Phi_n \}_{n \in \mathbb{N}} = \{ \Pi_x(x_n - \gamma H^{-1} \nabla_n f(x_n), H_n) \}_{n \in \mathbb{N}} \) is bounded.

2. Consider any \( H \in \mathcal{H} \). Then, there is a subsequence \( \{ n_k \}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( H_{n_k} \to H \) as \( k \to \infty \). Also,
\[
\Phi_{n_k} = \Pi_x \left( x_{n_k} - \gamma H_{n_k}^{-1} \nabla_{n_k} f(x_{n_k}), H_{n_k} \right)
\]
Taking the limit as \( k \to \infty \), it follows from the continuity of \( \Pi_x \) that
\[
\lim_{k \to \infty} \Phi_{n_k} = \Pi_x \left( x - \gamma H^{-1} \nabla f(x), H \right)
\]
that is, \( \Pi_x \left( x - \gamma H^{-1} \nabla f(x), H \right) \in \mathcal{Y}_\Phi \).

Next, consider any \( z \in \mathcal{Y}_\Phi \). Then, there is a subsequence \( \{ n_k \}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( \Phi_{n_k} \to z \) as \( k \to \infty \). Since \( \{ H_{n_k} \}_{k \in \mathbb{N}} \subset \mathcal{P} \) and \( \mathcal{P} \) is compact, there is a further subsequence \( \{ m_k \}_{k \in \mathbb{N}} \subset \{ n_k \}_{k \in \mathbb{N}} \) such that \( H_{m_k} \to H \in \mathcal{H} \) as \( k \to \infty \). Also,
\[
\Phi_{m_k} = \Pi_x \left( x_{m_k} - \gamma H_{m_k}^{-1} \nabla_{m_k} f(x_{m_k}), H_{m_k} \right)
\]
Taking the limit as \( k \to \infty \), it follows from the continuity of \( \Pi_x \) that
\[
z = \Pi_x (x - \gamma H^{-1} \nabla f(x), H)
\]
3. Suppose that $x \notin \mathcal{S}$. Consider any $z \in \mathcal{Y}_\Phi$. It follows from the previous part that there is $H \in \mathcal{H}$ such that $z = \Pi_X(x - \gamma H^{-1}\nabla f(x), H)$. Then it follows from Properties P 7.5 and P 7.6 that

$$\nabla f(x)^T(z - x) \leq -\frac{\|z - x\|_H^2}{\gamma} \leq -\frac{\lambda_1}{\gamma} \|z - x\|_2^2 < 0$$

Thus, $\nabla f(x)^T(z - x) < 0$ for every $z \in \mathcal{Y}_\Phi$. Since $\mathcal{Y}_\Phi$ is nonempty and compact, and $\nabla f(x)^T(z - x)$ is continuous in $z$, it follows that

$$\sup_{z \in \mathcal{Y}_\Phi} \nabla f(x)^T(z - x) < 0$$

4. If $x \in \mathcal{S}$, then it follows from Property P 7.6 that $\Pi_X(x - \gamma H^{-1}\nabla f(x), H) = x$ for every $H \in \mathcal{H}$. Therefore, $\mathcal{Y}_\Phi = \{x\}$ and hence

$$\sup_{z \in \mathcal{Y}_\Phi} \nabla f(x)^T(z - x) = 0$$

Lemma 7.17 showed that for any $\alpha \in [0, 1]$, $\mathcal{Y}(\alpha) = \{x + \alpha(z - x) : z \in \mathcal{Y}_\Phi\}$. Therefore, if $x \in \mathcal{S}$, then for all $\alpha \in [0, 1]$,

$$\mathcal{Y}(\alpha) = \{x\}$$

\[\square\]

**Lemma 7.21.** Consider a sequence $\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X}$ such that $x_n \to x \in \mathcal{X}$ as $n \to \infty$. Consider a corresponding sequence $\{\Phi_n\}_{n \in \mathbb{N}}$ of target points defined in (160) for some $\gamma > 0$ and $\{H_n\}_{n \in \mathbb{N}} \subset \mathcal{P}$. Suppose that $f$ is continuously differentiable on a neighborhood of $\mathcal{X}$, and that $\{\hat{\nabla}_n f\}_{n \in \mathbb{N}}$ satisfies Assumption A 5.2. Let $\mathcal{S}$ denote the set of stationary points of $f$ on $\mathcal{X}$, as defined in (109). Let $\mathcal{H}$ denote the set of limit points of the sequence $\{H_n\}_{n \in \mathbb{N}}$. Then the following properties hold.

1. The sequence $\{\Phi_n\}_{n \in \mathbb{N}}$ of target points is bounded. Hence the set $\mathcal{Y}_\Phi$ of limit points of the sequence $\{\Phi_n\}_{n \in \mathbb{N}}$ is nonempty and compact.

2. The set $\mathcal{Y}_\Phi$ satisfies

$$\mathcal{Y}_\Phi \subset \{\hat{\Pi}_x(x, \gamma, -H^{-1}\nabla f(x), H) : H \in \mathcal{H}\}$$

3. If $x \notin \mathcal{S}$, then

$$\nabla f(x)^T(z - x) \leq -\frac{\eta_1 \lambda_1}{\gamma} \|z - x\|_2^2 < 0$$

$$\sup_{z \in \mathcal{Y}_\Phi} \nabla f(x)^T(z - x) < 0$$

100
4. If \( x \in \mathcal{S} \), then
\[
\mathcal{Y}_\Phi = \{ x \}
\]

Consequently, the sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs defined in (154) satisfy Properties P 7.21 and P 7.22. In addition, if \( x \in \mathcal{S} \), then for all \( \alpha \in [0, 1] \),
\[
\mathcal{Y}(\alpha) = \{ x \}
\]

Proof. 1. It follows from Lemma 7.12 that for any \( \gamma \in [0, \infty) \), the sequence \( \{ \Phi_n \}_{n \in \mathbb{N}} \in \{ \hat{\Pi}_x(x_n, \gamma, -H_n^{-1} \nabla f(x_n)) \} \) is bounded.

2. Consider any \( z \in \mathcal{Y}_\Phi \). Then there is a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( \Phi_{n_k} \to z \) as \( k \to \infty \). Since \( \{H_{n_k}\}_{k \in \mathbb{N}} \subset \mathcal{P} \) and \( \mathcal{P} \) is compact, there is a further subsequence \( \{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}} \) such that \( H_{m_k} \to H \in \mathcal{P} \). Also,
\[
\Phi_{m_k} \in \hat{\Pi}_x(x_{m_k}, \gamma, -H_{m_k}^{-1} \nabla f(x_{m_k}), H_{m_k})
\]

It follows from (129) and (130) that
\[
\frac{3\gamma}{\sqrt{1-\eta_1}} \left\| \Phi_{m_k} - \Pi_x \left( x_{m_k} - \gamma H_{m_k}^{-1} \nabla f(x_{m_k}), H_{m_k} \right) \right\|_{H_{m_k}} \left\| H_{m_k}^{-1} \nabla f(x_{m_k}) \right\|_{H_{m_k}} \leq \left( \frac{\gamma}{\alpha'} \right)^{\zeta/2-1} \left\| \Phi_{m_k} - x_{m_k} \right\|_{H_{m_k}}^2 \quad (161)
\]

and
\[
\left\| \Phi_{m_k} - \Pi_x \left( x_{m_k} - \gamma H_{m_k}^{-1} \nabla f(x_{m_k}), H_{m_k} \right) \right\|_{H_{m_k}} \leq \frac{3\gamma}{\sqrt{1-\eta_1}} \left( \frac{\gamma}{\alpha'} \right)^{\zeta/2-1} \left\| H_{m_k}^{-1} \nabla f(x_{m_k}) \right\|_{H_{m_k}} \quad (162)
\]

Taking the limit as \( k \to \infty \), it follows from the continuity of \( \Pi_x \) that
\[
\frac{3\gamma}{\sqrt{1-\eta_1}} \left\| z - \Pi_x \left( x - \gamma H^{-1} \nabla f(x), H \right) \right\|_H \left\| H^{-1} \nabla f(x) \right\|_H \leq \left( \frac{\gamma}{\alpha'} \right)^{\zeta/2-1} \left\| z - x \right\|_H^2
\]

and
\[
\left\| z - \Pi_x \left( x - \gamma H^{-1} \nabla f(x), H \right) \right\|_H \leq \frac{3\gamma}{\sqrt{1-\eta_1}} \left( \frac{\gamma}{\alpha'} \right)^{\zeta/2-1} \left\| H^{-1} \nabla f(x) \right\|_H
\]

Thus, for any \( z \in \mathcal{Y}_\Phi \), there is \( H \in \mathcal{H} \) such that \( z \in \hat{\Pi}_x(x, \gamma, -H^{-1} \nabla f(x), H) \).
3. Suppose that $x \in S$. Consider any $z \in \mathcal{Y}_\phi$. Then there is $H \in \mathcal{H}$ such that $z \in \hat{\Pi}_x(x, \gamma, -H^{-1}\nabla f(x), H)$.

Since $0 < \gamma \leq [(1 - \eta_1)/4]^{1/(\zeta - 2)} \bar{\alpha}'$, it follows from Corollary 7.8 and 114 that

$$\nabla f(x)^T (z - x) \leq \frac{-\eta_1}{\gamma} \|z - x\|_H^2 \leq \frac{-\eta_1 \lambda_1}{\gamma} \|z - x\|_2^2$$

It follows from (137) in Lemma 7.10 that $\|z - x\|_H > 0$ and hence $\nabla f(x)^T (z - x) < 0$. Since $\mathcal{Y}_\phi$ is nonempty and compact, and $\nabla f(x)^T (z - x)$ is continuous in $z$, it follows that

$$\sup_{z \in \mathcal{Y}_\phi} \nabla f(x)^T (z - x) < 0$$

4. Suppose that $x \in S$. Consider any $z \in \mathcal{Y}_\phi$, and let $H \in \mathcal{H}$ be such that $z \in \hat{\Pi}_x(x, \gamma, -H^{-1}\nabla f(x), H)$.

Since $\gamma < \bar{\alpha}'$, it follows from Part b of Lemma 7.10 that $\hat{\Pi}_x(x, \gamma, -H^{-1}\nabla f(x), H) = \{x\}$.

Thus $z = x$, and hence $\mathcal{Y}_\phi = \{x\}$ and

$$\sup_{z \in \mathcal{Y}_\phi} \nabla f(x)^T (z - x) = 0$$

Lemma 7.17 showed that for any $\alpha \in [0, 1]$, $\mathcal{Y}(\alpha) = \{x + \alpha(z - x) : z \in \mathcal{Y}_\phi\}$. Therefore, if $x \in S$, then for all $\alpha \in [0, 1]$,

$$\mathcal{Y}(\alpha) = \{x\} \quad \square$$

### 7.2 Step-Size Rules

Having discussed arc rules, in this section we consider a few widely used methods to choose step-sizes $\alpha_n$, and thereby to choose steps $h_n(\alpha_n)$ at each iteration $n$. Further, we analyze the algorithms obtained by the combination of the arc rules of the previous section with these step-size rules and prove their convergence using the framework of Section 6 when objective values and derivatives are approximated.

#### 7.2.1 The Armijo Rule

In this section we analyze a line-search algorithm based on the well-known Armijo step-size rule. First, we state the Armijo sufficient conditions in the case where $f(x)$ and $\nabla f(x)$ can be evaluated...
exactly. Choose constants $\bar{\alpha} > 0$, and $\eta, \tau \in (0, 1)$. At iteration $n$, given iterate $x_n$ and arc $h_n$, the step-size $\alpha_n$ is chosen to satisfy the following sufficient improvement condition:

$$ f(x_n) - f(h_n(\alpha_n)) \geq -\eta \nabla f(x_n)^T (h_n(\alpha_n) - x_n) > 0 $$ \hspace{1cm} (163) 

and the following sufficient step-size condition

$$ \alpha_n \geq \tau \bar{\alpha}_n $$ \hspace{1cm} (164) 

where $\bar{\alpha}_n$ satisfies either

$$ f(x_n) - f(h_n(\bar{\alpha}_n)) < -\eta \nabla f(x_n)^T (h_n(\bar{\alpha}_n) - x_n) $$ \hspace{1cm} (165) 

or $\bar{\alpha}_n \geq \bar{\alpha}$.

Any step-size $\alpha_n$ that satisfies the Armijo sufficient conditions is called an Armijo step-size.

The following lemma establishes the existence of an Armijo step-size whenever the arc generating function $h_n$ satisfies certain conditions and $x_n \in \mathcal{X} \setminus \mathcal{S}$.

**Lemma 7.22.** Suppose that $f$ is continuously differentiable at $x_n$. Suppose that the arc $h_n$ satisfies the following conditions.

$$ \lim_{\alpha \to 0} h_n(\alpha) = x_n $$ \hspace{1cm} (166) 

$$ \nabla f(x_n)^T [h_n(\alpha) - x_n] < 0 \quad \forall \quad \alpha \in (0, \bar{\alpha}] $$ \hspace{1cm} (167) 

$$ \limsup_{\alpha \to 0} \frac{\nabla f(x_n)^T [h_n(\alpha) - x_n]}{\|h_n(\alpha) - x_n\|_2} < 0 $$ \hspace{1cm} (168) 

Then, there exists $\alpha_n \in (0, \bar{\alpha}]$ such that the Armijo sufficient conditions (163) and (164) are satisfied.

**Proof.** Since $f$ is continuously differentiable at $x_n$,

$$ \lim_{y \to x_n} \frac{f(x_n) - f(y) - \nabla f(x_n)^T (x_n - y)}{\|y - x_n\|_2} = 0 $$

Thus it follows from (166) that

$$ \lim_{\alpha \to 0} \frac{f(x_n) - f(h_n(\alpha)) - \nabla f(x_n)^T [x_n - h_n(\alpha)]}{\|h_n(\alpha) - x_n\|_2} = 0 $$ \hspace{1cm} (169) 

It follows from (168) that there is $\varepsilon > 0$ and $\alpha_1 \in (0, \bar{\alpha}]$ such that for all $\alpha \in (0, \alpha_1)$,

$$ (1 - \eta) \nabla f(x_n)^T \frac{h_n(\alpha) - x_n}{\|h_n(\alpha) - x_n\|_2} < -\varepsilon $$ \hspace{1cm} (170)
It follows from (169) and (170) that there is \( \alpha_2 \in (0, \alpha_1) \) such that for all \( \alpha \in (0, \alpha_2) \),

\[
\frac{f(x_n) - f(h_n(\alpha)) - \nabla f(x_n)^T [x_n - h_n(\alpha)]}{\|h_n(\alpha) - x_n\|_2} \geq -\varepsilon > (1 - \eta) \frac{\nabla f(x_n)^T h_n(\alpha) - x_n}{\|h_n(\alpha) - x_n\|_2}
\]

\[
\Rightarrow f(x_n) - f(h_n(\alpha)) > -\eta \nabla f(x_n)^T [h_n(\alpha) - x_n]
\]

It also follows from (170) that for all \( \alpha \in (0, \alpha_1) \),

\[
-\eta \nabla f(x_n)^T [h_n(\alpha) - x_n] > \frac{\eta \varepsilon}{1 - \eta} \|h_n(\alpha) - x_n\|_2 > 0
\]

Thus, for all \( \alpha \in (0, \alpha_2) \),

\[
f(x_n) - f(h_n(\alpha)) > -\eta \nabla f(x_n)^T [h_n(\alpha) - x_n] > 0
\]

(171)

Let

\[
m_n := \min \left\{ m \in \{0, 1, \ldots\} : f(x_n) - f(h_n(\tau^m \bar{\alpha})) \geq -\eta \nabla f(x_n)^T [h_n(\tau^m \bar{\alpha}) - x_n] \right\}
\]

Since \( \tau \in (0, 1) \), there is \( m \in \mathbb{N} \) such that \( 0 < \tau^m \bar{\alpha} < \alpha_2 \), and hence \( m_n < \infty \). Then

\[
f(x_n) - f(h_n(\tau^{m_n} \bar{\alpha})) \geq -\eta \nabla f(x_n)^T [h_n(\tau^{m_n} \bar{\alpha}) - x_n] > 0
\]

where the last inequality follows from (167). That is, the Armijo sufficient conditions (163) and (164) are satisfied with \( \bar{\alpha}_n = \tau^{m_n-1} \bar{\alpha} \) and \( \alpha_n = \tau^{m_n} \bar{\alpha} \).

The proof of Lemma 7.22 not only guarantees the existence of an Armijo step-size but also provides a method to find it whenever \( x_n \in \mathcal{X} \setminus \mathcal{S} \).

Next we consider the case where \( f \) and \( \nabla f \) are approximated at iteration \( n \) by \( f_n \) and \( \hat{\nabla}_n f \) respectively. We adapt the Armijo step-size rule for this setting.

Under conditions (166 through (168), Lemma 7.22 guaranteed the existence of \( m \in \mathbb{N} \) such that

\[
f(x_n) - f(h_n(\tau^m \bar{\alpha})) \geq -\eta \nabla f(x_n)^T [h_n(\tau^m \bar{\alpha}) - x_n] > 0,
\]

and thereby the existence of an Armijo step-size whenever \( x_n \in \mathcal{X} \setminus \mathcal{S} \). However, if the approximations \( f_n \) and \( \hat{\nabla}_n f \) are used instead of \( f \) and \( \nabla f \), then even if \( x_n \in \mathcal{X} \setminus \mathcal{S} \), there may be no \( m \in \mathbb{N} \) such that \( f_n(x_n) - f_n(h_n(\tau^m \bar{\alpha})) \geq -\eta \hat{\nabla}_n f(x_n)^T [h_n(\tau^m \bar{\alpha}) - x_n] > 0 \).

In order to force finite termination of the search for an Armijo step-size at every iteration \( n \), we introduce a sequence \( \{u_n\}_{n \in \mathbb{N}} \subset \mathbb{N} \), such that \( u_n \to \infty \) as \( n \to \infty \). At each iteration \( n \), we search
for the smallest integer \( m \in \{0, \ldots, u_n\} \) that satisfies

\[
  f_n(x_n) - f_n(h_n(\tau^m \bar{\alpha})) \geq -\eta \nabla_n f(x_n)^T [h_n(\tau^m \bar{\alpha}) - x_n] > 0
\]

If such an integer \( m \) is found, then we set \( x_{n+1} = h_n(\tau^m \bar{\alpha}) \). Otherwise, we let \( x_{n+1} = x_n \) and start the next iteration \( n + 1 \) with refined function and gradient estimates \( f_{n+1} \) and \( \nabla f_{n+1} \). The sequence \( \{u_n\}_{n \in \mathbb{N}} \) may depend on data obtained during the course of the algorithm, and does not have to be chosen before the start of the algorithm, as long as \( u_n \to \infty \) as \( n \to \infty \).

**Algorithm 7.1.** Choose constants \( \bar{\alpha} > 0 \), and \( \eta, \tau \in (0, 1) \). Choose an initial feasible point \( x_0 \in \mathcal{X} \). For \( n = 0, 1, \ldots \) until a chosen stopping criterion is satisfied, do the following steps:

**Step 1:** Choose an arc \( h_n : [0, \bar{\alpha}] \to \mathcal{X} \) that satisfies the conditions of Lemma 7.22.

**Step 2:** If there is \( m \in \{0, 1, \ldots, u_n\} \) such that \( f_n(x_n) - f_n(h_n(\tau^m \bar{\alpha})) \geq -\eta \nabla_n f(x_n)^T (h_n(\tau^m \bar{\alpha}) - x_n) \), then let

\[
  m_n := \min \{ m \in \{0, 1, \ldots, u_n\} : f_n(x_n) - f_n(h_n(\tau^m \bar{\alpha})) \geq -\eta \nabla_n f(x_n)^T (h_n(\tau^m \bar{\alpha}) - x_n) \}
\]

and set \( x_{n+1} = h_n(\tau^{m_n} \bar{\alpha}) \); otherwise set \( x_{n+1} = x_n \).

Next we analyze the convergence of Algorithm 7.1. We use the framework developed in Section 6. Specifically, we establish sufficient conditions for the sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 7.1 to satisfy the asymptotic descent property in Definition 6.1, and then we use Theorem 6.9 to show that \( \{x_n\}_{n \in \mathbb{N}} \) satisfies (112).

As in the case of the exact Armijo algorithm, we impose some conditions on the arc rule to obtain a sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 7.1 that satisfies the asymptotic descent property. Consider any subsequence \( \{x_{n_k}\}_{k \in \mathbb{N}} \) of the sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 7.1, such that \( x_{n_k} \to x \) for some \( x \in \mathcal{X} \). For any \( \alpha \in [0, \bar{\alpha}] \), let \( \mathcal{Y}(\alpha) \) denote the set of limit points of the corresponding sequence \( \{h_{n_k}(\alpha)\}_{k \in \mathbb{N}} \). (Thus \( \mathcal{Y}(\alpha) \) depends on \( \{x_{n_k}\}_{k \in \mathbb{N}} \); however, this dependence is not indicated in the notation and should be clear from the context.)

**A 7.2.** For any \( \alpha \in [0, \bar{\alpha}] \), \( \mathcal{Y}(\alpha) \) is nonempty.

**A 7.3.** For any \( \{\alpha_{n_k}\}_{k \in \mathbb{N}} \subset [0, \bar{\alpha}] \) such that \( \alpha_{n_k} \to 0 \) as \( k \to \infty \), it holds that

\[
  \lim_{k \to \infty} h_{n_k}(\alpha_{n_k}) = x
\]  

(172)
A 7.4. If \( x \notin S \), then for any \( \alpha \in (0, \bar{\alpha}] \) and any \( y \in \mathcal{Y}(\alpha) \),

\[
\|y - x\|_2 > 0
\]

\[
\nabla f(x)^T(y - x) < 0
\]

(173)

(174)

In addition,

\[
\limsup_{\alpha \downarrow 0} \sup_{y \in \mathcal{Y}(\alpha)} \nabla f(x)^T \frac{y - x}{\|y - x\|_2} < 0
\]

(175)

A 7.5. If \( x \in S \), then for any \( \alpha \in [0, \bar{\alpha}] \) and any \( y \in \mathcal{Y}(\alpha) \),

\[
\nabla f(x)^T(y - x) \leq 0
\]

(176)

Before we show that the sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 7.1 satisfies the asymptotic descent property, we first verify that the arc rules described in the previous section satisfy Assumptions A7.2 – A7.5.

**Example 7.4.** Consider the scaled gradient projection arc rule defined in (124). That is, at iteration \( n \), given \( x_n \in X \) and \( H_n \in \mathcal{P} \),

\[
h_n(\alpha) := \Pi_X(x_n - \alpha H_n^{-1}\hat{\nabla}_n f(x_n), H_n)
\]

for each \( \alpha \in [0, \infty) \). Property P 7.8 verifies Assumption A 7.2; Property P 7.9 verifies Assumption A 7.3; Property P 7.11 verifies Assumption A 7.4; and Property P 7.12 verifies Assumption A 7.5.

**Example 7.5.** Consider the scaled gradient inexact projection arc rule defined in (142). That is, at iteration \( n \), given \( x_n \in X \) and \( H_n \in \mathcal{P} \),

\[
h_n(\alpha) \in \hat{\Pi}_X(x_n, \alpha, -H_n^{-1}\hat{\nabla}_n f(x_n), H_n)
\]

(176)

for each \( \alpha \in [0, \bar{\alpha}] \). Property P 7.13 verifies Assumption A 7.2; Property P 7.14 verifies Assumption A 7.3; Property P 7.16 verifies Assumption A 7.4; and Property P 7.17 verifies Assumption A 7.5.

**Example 7.6.** Consider the feasible direction arc rule defined in (154). That is, at iteration \( n \), given \( x_n, \Phi_n \in X \),

\[
h_n(\alpha) := x_n + \alpha(\Phi_n - x_n)
\]

for each \( \alpha \in [0, 1] \). In Section 7.1.3 we described three target point rules (Examples 7.1 through 7.3) for choosing \( \Phi_n \). We also showed in Lemmas 7.19 through 7.21 that each of these target point rules gives an arc rule that satisfies Properties P 7.18 – P 7.22. Property P 7.18 verifies Assumption A 7.2; Property P 7.19 verifies Assumption A 7.3; Property P 7.21 verifies Assumption A 7.4; and Property P 7.22 verifies Assumption A 7.5.
Next we show that Algorithm 7.1 satisfies the asymptotic descent property of Definition 6.1. First, we establish the following result.

**Lemma 7.23.** Suppose that $f$ is continuously differentiable on a neighborhood of $\mathcal{X}$, and that $\{f_n\}_{n \in \mathbb{N}}$ and $\{\hat{\nabla}_n f\}_{n \in \mathbb{N}}$ satisfy Assumption A 5.2. Let $\mathcal{S}$ denote the set of stationary points of $f$ on $\mathcal{X}$, as defined in (109). Consider the sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.1 with parameters $\bar{\alpha} > 0$, $\eta, \tau \in (0, 1)$, and sequence $\{u_n\}_{n \in \mathbb{N}}$ such that $u_n \to \infty$. Let $\{h_n\}_{n \in \mathbb{N}}$ be a corresponding sequence of arcs. Consider any subsequence $\{n_k\}_{k \in \mathbb{N}}$ such that $x_{n_k} \to x \in \mathcal{X} \setminus \mathcal{S}$ as $k \to \infty$. Suppose that the sequence $\{h_{n_k}\}_{k \in \mathbb{N}}$ of arcs satisfies Assumptions A 7.2, A 7.3, and A 7.4. Then, the following hold.

(a) There exists $m_x \in \mathbb{N}$ such that for all $y \in \mathcal{Y}(\tau^{m_x} \bar{\alpha})$,

$$f(x) - f(y) > -\eta \nabla f(x)^T(y - x) > 0$$

(b) Consider any $y \in \mathcal{Y}(\tau^{m_x} \bar{\alpha})$, and corresponding subsequence $\{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}}$ such that $h_{m_k}(\tau^{m_x} \bar{\alpha}) \to y$ as $k \to \infty$. Then there exists $N \in \mathbb{N}$ such that for all $k > N$, $m_x < u_{m_k}$ and

$$f_{m_k}(x_{m_k}) - f_{m_k}(h_{m_k}(\tau^{m_x} \bar{\alpha})) > -\eta \hat{\nabla}_{m_k} f(x_{m_k})^T(h_{m_k}(\tau^{m_x} \bar{\alpha}) - x_{m_k}) > 0$$

**Proof.** (a) Since $x \in \mathcal{X} \setminus \mathcal{S}$, it follows from (175) of Assumption A 7.4 that there is $\varepsilon > 0$ and $\alpha' > 0$ such that for all $\alpha \in (0, \alpha')$ and all $y \in \mathcal{Y}(\alpha)$,

$$\nabla f(x)^T \frac{y - x}{\|y - x\|_2} < \frac{-\varepsilon}{1 - \eta} \quad (177)$$

Since $f$ is continuously differentiable at $x$, it holds that

$$\lim_{y \to x} \frac{f(x) - f(y) - \nabla f(x)^T(x - y)}{\|y - x\|_2} = 0$$

Thus, there is $\delta > 0$ such that for all $y \in \mathbb{R}^l$ with $\|y - x\|_2 < \delta$,

$$-\varepsilon < \frac{f(x) - f(y) - \nabla f(x)^T(x - y)}{\|y - x\|_2} < \varepsilon \quad (178)$$

It follows from Assumption A 7.2 that $\mathcal{Y}(\alpha)$ is nonempty for each $\alpha \in [0, \bar{\alpha}]$, and hence it follows from Assumption A 7.3 and Lemma 7.1 that

$$\lim_{\alpha \to 0} \sup_{y \in \mathcal{Y}(\alpha)} \|y - x\|_2 = 0$$
Thus, there is \( \alpha'' \in (0, \alpha'] \) such that for all \( \alpha \in (0, \alpha'') \) and all \( y \in \mathcal{Y}(\alpha) \), \( \|y - x\|_2 < \delta \). Then it follows from (177) and (178) that for all \( \alpha \in (0, \alpha'') \) and all \( y \in \mathcal{Y}(\alpha) \),

\[
(1 - \eta) \nabla f(x)^T \frac{y - x}{\|y - x\|_2} < -\varepsilon < \frac{f(x) - f(y) - \nabla f(x)^T (x - y)}{\|y - x\|_2} \Rightarrow f(x) - f(y) > -\eta \nabla f(x)^T (y - x)
\]

Also, it follows from (173) and (177) that

\[
-\eta \nabla f(x)^T (y - x) > \frac{\eta \varepsilon}{1 - \eta} \|y - x\|_2 > 0
\]

Since \( \tau \in (0, 1) \), there is \( m_x \in \mathbb{N} \) such that \( \tau^{m_x} \bar{\alpha} \in (0, \alpha'') \), and hence, for all \( y \in \mathcal{Y}(\tau^{m_x} \bar{\alpha}) \),

\[
f(x) - f(y) > -\eta \nabla f(x)^T (y - x) > 0
\]

(b) It follows from \( f(x) - f(y) > -\eta \nabla f(x)^T (y - x) > 0 \) that there exists \( \varepsilon > 0 \) such that \( f(x) - f(y) > -\eta \nabla f(x)^T (y - x) + 2\varepsilon \) and \( -\eta \nabla f(x)^T (y - x) > \varepsilon \). It follows from Assumption A 5.2 that \( f_{m_k}(x_{m_k}) \rightarrow f(x) \), \( f_{m_k}(h_{m_k}(\tau^{m_x} \bar{\alpha})) \rightarrow f(y) \), and \( \hat{\nabla}_{m_k} f(x_{m_k}) \rightarrow \nabla f(x) \) as \( k \rightarrow \infty \). Therefore, there is \( N_1 \in \mathbb{N} \) such that for all \( k > N_1 \),

\[
f_{m_k}(x_{m_k}) - f_{m_k}(h_{m_k}(\tau^{m_x} \bar{\alpha})) > f(x) - f(y) - \varepsilon
\]

and \( N_2 \in \mathbb{N} \) such that for all \( k > N_2 \),

\[
-\eta \nabla f(x)^T (y - x) - \varepsilon < -\eta \hat{\nabla}_{m_k} f(x_{m_k})^T (h_{m_k}(\tau^{m_x} \bar{\alpha}) - x_{m_k}) < -\eta \nabla f(x)^T (y - x) + \varepsilon
\]

Also, there is \( N_3 \in \mathbb{N} \) such that for all \( k > N_3 \), \( m_x < u_{m_k} \). Thus, for all \( k > N := \max(N_1, N_2, N_3) \), it holds that \( m_x < u_{m_k} \) and

\[
f_{m_k}(x_{m_k}) - f_{m_k}(h_{m_k}(\tau^{m_x} \bar{\alpha})) > f(x) - f(y) - \varepsilon
\]

\[
> -\eta \nabla f(x)^T (y - x) + \varepsilon
\]

\[
> -\eta \hat{\nabla}_{m_k} f(x_{m_k})^T (h_{m_k}(\tau^{m_x} \bar{\alpha}) - x_{m_k})
\]

\[
> -\eta \nabla f(x)^T (y - x) - \varepsilon > 0
\]

Next we show that Algorithm 7.1 satisfies the asymptotic descent property.
Thus, for each $X$ and $\{x \in Y \text{ and } m_k \leq x \}$ such that for all $h_k \rightarrow 0$, $h_k \rightarrow \infty$. Let $\{h_n\}_{n \in \mathbb{N}}$ be a corresponding sequence of arcs. Suppose that the sequence $\{h_n\}_{n \in \mathbb{N}}$ of arcs satisfies Assumptions A7.2 – A7.5. Then the sequence $\{x_n\}_{n \in \mathbb{N}}$ satisfies the asymptotic descent property in Definition 6.1.

**Proof.** We show that the sufficient conditions of Lemma 6.5 are satisfied. Consider any subsequence $\{x_n\}_{n \in \mathbb{N}}$ such that $x_{n_k} \rightarrow \bar{x} \in X$ and $x_{n_k+1} \rightarrow \bar{x} \in X$ as $k \rightarrow \infty$.

Suppose that $\bar{x} \notin S$. It follows from Lemma 7.23 that there is $m_{\bar{x}}, N \in \mathbb{N}$ and a further subsequence $\{m_k\}_{k \in \mathbb{N}} \subseteq \{n_k\}_{k \in \mathbb{N}}$ such that for all $k > N$, $m_{\bar{x}} < u_{m_{\bar{x}}}$ and

$$f_{m_k}(x_{m_k}) - f_{m_k}(h_{m_k}(\tau^{m_{\bar{x}}}) > -\eta\nabla f(x_{m_k})^T(h_{m_k}(\tau^{m_{\bar{x}}}) - x_{m_k}) > 0$$

Thus, for each $k > N$, Step 2 in Algorithm 7.1 finds a step-size $\alpha_{m_k} = \tau^{m_{\bar{x}}} \alpha$, where $m_{m_{\bar{x}}} := \min \{m \in \{0, 1, \ldots, u_{m_{\bar{x}}} : f_{m_k}(x_{m_k}) - f_{m_k}(h_{m_k}(\tau^{m_{\bar{x}}})) \geq -\eta\nabla f(x_{m_k})^T[h_{m_k}(\tau^{m_{\bar{x}}}) - x_{m_k}]\}

\leq m_{\bar{x}} \leq u_{m_{\bar{x}}}

Thus, $\{m_{m_{\bar{x}}}\}_{k \in \mathbb{N}}$ is a bounded sequence and hence there is a further subsequence $\{j_k\}_{k \in \mathbb{N}} \subseteq \{m_k\}_{k \in \mathbb{N}}$ such that for all $k$, $m_{j_k} = m \leq m_{\bar{x}}$. Hence, for all $k$,

$$x_{j_k+1} = h_{j_k}(\tau^m \alpha) \quad (179)$$

and

$$f_{j_k}(x_{j_k}) - f_{j_k}(x_{j_k+1}) \geq -\eta\nabla f(x_{j_k})^T(x_{j_k+1} - x_{j_k}) \quad (180)$$

Taking the limit as $k \rightarrow \infty$, it follows that $f(\bar{x}) - f(\bar{x}) \geq -\eta\nabla f(\bar{x})^T(\bar{x} - \bar{x})$. Since $\bar{x} \notin S$ and $\bar{x} \in S(\tau^m \alpha)$, it follows from Assumption A 7.4 that $\nabla f(\bar{x})^T(\bar{x} - \bar{x}) < 0$. Thus,

$$f(\bar{x}) - f(\bar{x}) \geq -\eta\nabla f(\bar{x})^T(\bar{x} - \bar{x}) > 0$$

Suppose that $\bar{x} \in S$. For each $k \in \mathbb{N}$, if there is $m \in \{0, 1, \ldots, u_{m_{\bar{x}}} \}$ such that $f_{m_k}(x_{m_k}) - f_{m_k}(h_{m_k}(\tau^{m_{\bar{x}}})) \geq -\eta\nabla f(x_{m_k})^T[h_{m_k}(\tau^{m_{\bar{x}}}) - x_{m_k}]$, then Step 2 in Algorithm 7.1 sets $m_{m_{\bar{x}}} := \min \{m \in \{0, 1, \ldots, u_{m_{\bar{x}}} : f_{m_k}(x_{m_k}) - f_{m_k}(h_{m_k}(\tau^{m_{\bar{x}}})) \geq -\eta\nabla f(x_{m_k})^T[h_{m_k}(\tau^{m_{\bar{x}}}) - x_{m_k}]\}$

and $x_{n_k+1} = h_{n_k}(\tau^{m_{n_k}} \alpha)$; otherwise $x_{n_k+1} = x_{n_k}$. Accordingly, the following two cases arise.
Case 1: There is a further subsequence \( \{ m_k \}_{k \in \mathbb{N}} \subset \{ n_k \}_{k \in \mathbb{N}} \), such that for every \( k \in \mathbb{N} \), \( x_{m_k+1} = x_{m_k} \). Then \( x_{m_k+1} \to \bar{x} \). We assumed that \( x_{n_k+1} \to \bar{x} \). Therefore, \( \bar{x} = \bar{x} \), which implies that \( f(\bar{x}) = f(\bar{x}) \).

Case 2: There is \( N \in \mathbb{N} \) such that for all \( k \geq N \), \( x_{n_k+1} = h_{n_k}(\tau^{m_n} \bar{\alpha}) \) for some nonnegative integer \( m_n \leq u_n \). Consider the following two cases concerning the sequence \( \{ m_n \}_{k \geq N} \):

(a) There is \( M \in \mathbb{N} \) such that \( m_n \in \{ 0, 1, \ldots, M \} \) for all \( k \geq N \). Thus, there is \( m \in \{ 0, 1, \ldots, M \} \) and a further subsequence \( \{ m_k \}_{k \in \mathbb{N}} \subset \{ n_k \}_{k \in \mathbb{N}} \) such that \( m_{mk} = m \) for all \( k \in \mathbb{N} \). Thus, for all \( k \in \mathbb{N} \),

\[
x_{m_k+1} = h_{m_k}(\tau^m \bar{\alpha})
\]

\[
f_{m_k}(x_{m_k}) - f_{m_k}(x_{m_k+1}) \geq -\eta \nabla f_{m_k}(x_{m_k})^T(x_{m_k+1} - x_{m_k})
\]

Taking the limit as \( k \to \infty \), it follows that \( f(\bar{x}) - f(\bar{x}) \geq -\eta \nabla f(\bar{x})^T(\bar{x} - \bar{x}) \). Since \( \bar{x} \in S \) and \( \bar{x} \in \mathcal{Y}(\tau^m \bar{\alpha}) \), it follows from Assumption A 7.5 that \( \nabla f(\bar{x})^T(\bar{x} - \bar{x}) \leq 0 \). Hence,

\[
f(\bar{x}) - f(\bar{x}) \geq -\eta \nabla f(\bar{x})^T(\bar{x} - \bar{x}) \geq 0
\]

(b) Thus, there is a further subsequence \( \{ m_k \}_{k \in \mathbb{N}} \subset \{ n_k \}_{k \in \mathbb{N}} \) such that \( m_{mk} \to \infty \) as \( k \to \infty \). Then \( \tau^{m_{mk}} \to 0 \) and \( \alpha_{mk} = \tau^{m_{mk}} \bar{\alpha} \to 0 \) as \( k \to \infty \). Therefore, it follows from Assumption A 7.3 that

\[
\lim_{k \to \infty} h_{m_k}(\alpha_{mk}) = \bar{x}
\]

We assumed that \( x_{n_k+1} = h_{n_k}(\tau^{m_n} \bar{\alpha}) \to \bar{x} \). Thus \( \bar{x} = \bar{x} \) and hence \( f(\bar{x}) = f(\bar{x}) \).

We have thus shown that any sequence \( \{ x_n \}_{n \in \mathbb{N}} \) generated by Algorithm 7.1 satisfies the asymptotic descent property. \( \square \)

Finally, we can state the convergence theorem for our algorithm

**Theorem 7.25.** Suppose that the assumptions of Theorem 7.24 hold. Further, suppose that the set \( X \) is compact. Then, any sequence \( \{ x_n \}_{n \in \mathbb{N}} \) generated by Algorithm 7.1 satisfies

\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in S} f(x)
\]

In addition, if Assumptions A 6.5 and A 6.6 are satisfied, then

\[
\lim_{n \to \infty} d(x_n, S) = 0
\]
Proof. We assumed that $\mathcal{X}$ is compact and $f$ is continuously differentiable on a neighborhood of $\mathcal{X}$. Thus, Assumptions A 6.1 and A 6.9 are satisfied. Theorem 7.24 established that any sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.1 satisfies the asymptotic descent property. Thus, it follows from Theorem 6.9 that
\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in \mathcal{F}} f(x)
\]
and, if Assumptions A 6.5 and A 6.6 are satisfied, then
\[
\lim_{n \to \infty} d(x_n, \mathcal{F}) = 0
\]

### 7.2.2 The Constant Step Size Rule

In this section we analyze a line-search algorithm based on the constant step-size rule. In this section we make the following assumption:

**A 7.6.** The function $f$ is Lipschitz continuously differentiable on a neighborhood of $\mathcal{X}$ with Lipschitz constant $L < \infty$. That is, $\nabla f$ satisfies
\[
\|\nabla f(y) - \nabla f(x)\|_2 \leq L \|y - x\|_2
\]  
(181)
for all $x, y \in \mathcal{X}$.

It can be shown (e.g., see ?) that if Assumption A 7.6 holds, then
\[
f(y) - f(x) \leq \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2
\]  
(182)
for all $x, y \in \mathcal{X}$.

Next we state an algorithm that uses the constant step-size rule.

**Algorithm 7.2.** Choose constant $\alpha > 0$. Choose an initial feasible point $x_0 \in \mathcal{X}$. For $n = 0, 1, \ldots$ until a chosen stopping criterion is satisfied, do the following steps:

**Step 1:** Choose an arc $h_n : [0, \bar{\alpha}] \mapsto \mathcal{X}$.

**Step 2:** Set $x_{n+1} = h_n(\alpha)$. 

111
Algorithm 7.2 describes an algorithm that uses the constant step-size rule irrespective of whether $f$ and $\nabla f$ are evaluated exactly or not. However, the choice of arcs $h_n$ change depending on whether $f$ and $\nabla f$ are evaluated exactly or not. If $f$ and $\nabla f$ are evaluated exactly and the scaled gradient projection arc rule in (123) is used, it can be shown that the sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.2 converges if $\alpha < 2/L$ (e.g., see ?). In this section, we consider the case when $f$ and $\nabla f$ are approximated by $f_n$ and $\hat{\nabla}_n f$ respectively. We show that the sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.2 satisfies the asymptotic descent property. First, we state the conditions that we impose on the arc rule.

Consider any subsequence $\{x_{n_k}\}_{k \in \mathbb{N}}$ of the sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.2, such that $x_{n_k} \to x$ for some $x \in \mathcal{X}$. Let $\mathcal{Y}(\alpha)$ denote the set of limit points of the corresponding sequence $\{h_{n_k}(\alpha)\}_{k \in \mathbb{N}}$.

A 7.7. The set $\mathcal{Y}(\alpha)$ is nonempty.

A 7.8. If $x \notin S$, then for all $y \in \mathcal{Y}(\alpha)$,

$$\|y - x\|_2 > 0$$

Also, there exists a constant $\beta > 0$ (independent of $\{x_n\}_{n \in \mathbb{N}}$, subsequence $\{x_{n_k}\}_{k \in \mathbb{N}}$ and limit point $x \in \mathcal{X}$), such that for all $y \in \mathcal{Y}(\alpha)$,

$$\nabla f(x)^T(y - x) \leq -\frac{\beta}{\alpha}\|y - x\|^2_2$$

A 7.9. If $x \in S$, then

$$\mathcal{Y}(\alpha) = \{x\}$$

Next we show that some arc rules satisfy the assumptions stated above.

(a) Consider the scaled gradient projection arc rule defined in (124). That is, at iteration $n$, given $x_n \in \mathcal{X}$ and $H_n \in \mathcal{P}$,

$$h_n(\alpha) := \Pi_{\mathcal{X}}(x_n - \alpha H_n^{-1} \hat{\nabla}_n f(x_n), H_n)$$

for each $\alpha \in (0, \infty)$. It follows from Lemma 7.5 that Assumptions A7.7 – A7.9 are satisfied. Specifically, Properties P7.8, P7.11, and P7.12 respectively show that Assumptions A7.7, A7.8 (with $\beta = \lambda_1$), and A7.9 are satisfied.
(b) Consider the scaled gradient inexact projection arc rule defined in (142). That is, at iteration
\( n \), given \( x_n \in \mathcal{X} \) and \( H_n \in \mathcal{P} \),
\[
h_n(\alpha) \in \Pi_{\mathcal{X}}(x_n, \alpha, -H_n^{-1}\nabla f(x_n), H_n)
\]
for each \( \alpha \in [0, \alpha] \). It follows from Lemma 7.14 that Assumptions A7.7 – A7.9 are satisfied. Specifically, Properties P7.13, P7.16, and P7.17 respectively show that Assumptions A7.7, A7.8 (with \( \beta = \eta_1 \lambda_1 \)), and A7.9 are satisfied for all \( \alpha \leq ((1 - \eta_1)/4)^{1/(\zeta-2)} \alpha' \).

c) Consider the feasible direction arc rule defined in (154). That is, at iteration \( n \), given \( \Phi_n \in \mathcal{X} \),
\[
h_n(\alpha) := x_n + \alpha(\Phi_n - x_n)
\]
for each \( \alpha \in [0, 1] \). In Section 7.1.3 we described three target point rules (Examples 7.1 through 7.3) for choosing \( \Phi_n \). First consider the scaled gradient projection target point rule given in (158). That is, at iteration \( n \), given \( \gamma > 0 \), \( x_n \in \mathcal{X} \), \( \nabla f(x_n) \in \mathbb{R}^l \), and \( H_n \in \mathbb{R}^{l \times l} \), choose
\[
\Phi_n := \Pi_{\mathcal{X}}(x_n - \gamma H_n^{-1}\nabla f(x_n), H_n)
\]
Lemmas 7.18 and 7.20 show that Assumptions A 7.7 and A 7.9 are satisfied. Lemma 7.17 showed that for any \( \alpha \in [0, 1] \), \( \mathcal{Y}(\alpha) = \{ x + \alpha(z - x) : z \in \mathcal{Y}_k \} \). Thus, for any \( y \in \mathcal{Y}(\alpha) \), there is \( z \in \mathcal{Y}_k \) such that \( y = x + \alpha(z - x) \), and hence \( \nabla f(x)^T(y - x) = \alpha \nabla f(x)^T(z - x) \) and \( \|y - x\|_2 = \alpha \|z - x\|_2 \). If \( x \notin \mathcal{S} \), then it follows from Lemma 7.20 that \( \nabla f(x)^T(z - x) \leq -\lambda_1 \|z - x\|_2^2 / \gamma < 0 \), and thus, for any \( \alpha \in (0, 1] \),
\[
\nabla f(x)^T(y - x) = \alpha \nabla f(x)^T(z - x) \leq -\frac{\alpha \lambda_1}{\gamma} \|z - x\|_2^2 = -\frac{\lambda_1}{\gamma \alpha} \|y - x\|_2^2 < 0
\]
Hence Assumption A 7.8 holds with \( \beta = \lambda_1 / \gamma \).

Next consider the scaled gradient inexact projection target point rule given in (160). That is, at iteration \( n \), given \( \gamma > 0 \), \( x_n \in \mathcal{X} \), \( \nabla f(x_n) \in \mathbb{R}^l \), and \( H_n \in \mathbb{R}^{l \times l} \), choose
\[
\Phi_n \in \hat{\Pi}_{\mathcal{X}}(x_n, \gamma, -H_n^{-1}\nabla f(x_n), H_n)
\]
Lemmas 7.18 and 7.21 show that Assumptions A 7.7 and A 7.9 are satisfied. Lemma 7.17 showed that for any \( \alpha \in [0, 1] \), \( \mathcal{Y}(\alpha) = \{ x + \alpha(z - x) : z \in \mathcal{Y}_k \} \). Thus, for any \( y \in \mathcal{Y}(\alpha) \), there is \( z \in \mathcal{Y}_k \) such that \( y = x + \alpha(z - x) \), and hence \( \nabla f(x)^T(y - x) = \alpha \nabla f(x)^T(z - x) \)
and \(\|y - x\|_2 = \alpha \|z - x\|_2\). If \(x \notin \mathcal{S}\), then it follows from Lemma 7.21 that \(\nabla f(x)^T (z - x) \leq -\eta_1 \lambda_1 \|z - x\|_2^2 / \gamma < 0\), and thus, for any \(\alpha \in (0, 1]\),
\[
\nabla f(x)^T (y - x) = \alpha \nabla f(x)^T (z - x) \leq -\frac{\alpha \eta_1 \lambda_1}{\gamma \alpha} \|z - x\|_2^2 = -\frac{\eta_1 \lambda_1}{\gamma} \|y - x\|_2^2 < 0
\]

Hence Assumption A 7.8 holds with \(\beta = \eta_1 \lambda_1 / \gamma\).

**Theorem 7.26.** Suppose that \(f\) is Lipschitz continuously differentiable with Lipschitz constant \(L\) on a neighborhood of \(\mathcal{X}\), and that \(\{f_n\}_{n \in \mathbb{N}}\) and \(\{\nabla_n f\}_{n \in \mathbb{N}}\) satisfy Assumption A 5.2. Let \(\mathcal{S}\) denote the set of stationary points of \(f\) on \(\mathcal{X}\), as defined in (109). Consider the sequence \(\{x_n\}_{n \in \mathbb{N}}\) generated by Algorithm 7.2 with parameter \(\alpha \in (0, \beta / L)\), where \(\beta\) satisfies Assumption A 7.8. Let \(\{h_n\}_{n \in \mathbb{N}}\) be a corresponding sequence of arcs. Suppose that the sequence \(\{h_n\}_{n \in \mathbb{N}}\) of arcs satisfies Assumptions A7.7 – A7.9. Then the sequence \(\{x_n\}_{n \in \mathbb{N}}\) satisfies the asymptotic descent property in Definition 6.1.

**Proof.** We show that the sufficient conditions of Lemma 6.5 are satisfied. Consider any subsequence \(\{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N}\) such that \(x_{n_k} \to \bar{x} \in \mathcal{X}\) and \(x_{n_k+1} \to \bar{x} \in \mathcal{X}\) as \(k \to \infty\). Note that \(x_{n+1} = h_n(\alpha)\) for all \(n\), and thus \(\mathcal{Y}(\alpha) = \{\bar{x}\}\).

Suppose that \(\bar{x} \notin \mathcal{S}\). It follows from Assumption A 7.8 that
\[
\|\bar{x} - \bar{x}\|_2 > 0
\]
\[
\nabla f(\bar{x})^T (\bar{x} - \bar{x}) \leq -\frac{\beta}{\alpha} \|\bar{x} - \bar{x}\|_2^2
\]
(187)

Also, since \(f\) is Lipschitz continuously differentiable on \(\mathcal{X}\) with Lipschitz constant \(L\) and \(\bar{x}, \bar{x} \in \mathcal{X}\), it follows from (182) that
\[
f(\bar{x}) - f(\bar{x}) \leq \nabla f(\bar{x})^T (\bar{x} - \bar{x}) + \frac{L}{2} \|\bar{x} - \bar{x}\|_2^2
\]
Combining the above inequality with (187) it follows that
\[
f(\bar{x}) - f(\bar{x}) \leq \left(\frac{L}{2} - \frac{\beta}{\alpha}\right) \|\bar{x} - \bar{x}\|_2^2
\]
Since \(\alpha < 2\beta / L\) and \(\|\bar{x} - \bar{x}\|_2 > 0\), it follows that \(f(\bar{x}) < f(\bar{x})\).

Suppose that \(\bar{x} \in \mathcal{S}\). It follows from Assumption A 7.9 that \(\mathcal{Y}(\alpha) = \{\bar{x}\}\), which means that \(\bar{x} = \bar{x}\) and \(f(\bar{x}) = f(\bar{x})\).

Therefore, the sequence \(\{x_n\}_{n \in \mathbb{N}}\) satisfies the asymptotic descent property.
Theorem 7.27. Suppose that the assumptions of Theorem 7.26 hold. Further, suppose that the set $\mathcal{X}$ is compact. Then, any sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.2 satisfies
\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in \mathcal{S}} f(x)
\]
In addition, if Assumptions A 6.5 and A 6.6 are satisfied, then
\[
\lim_{n \to \infty} d(x_n, \mathcal{S}) = 0
\]

Proof. We assumed that $\mathcal{X}$ is compact and $f$ is continuously differentiable on a neighborhood of $\mathcal{X}$. Thus, Assumptions A 6.1 and A 6.9 are satisfied. Theorem 7.26 established that any sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.2 satisfies the asymptotic descent property. Thus, it follows from Theorem 6.9 that
\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in \mathcal{S}} f(x)
\]
and, if Assumptions A 6.5 and A 6.6 are satisfied, then
\[
\lim_{n \to \infty} d(x_n, \mathcal{S}) = 0
\]

7.2.3 The Diminishing Step-Size Rule

In this section we analyze a line-search algorithm based on the diminishing step-size rule. In this section we make Assumption A 7.6 and the following assumption:

A 7.10. The function $f$ is bounded below on $\mathcal{X}$, i.e.,
\[
\inf_{x \in \mathcal{X}} f(x) > -\infty
\]

78

First, we state an algorithm that uses the diminishing step-size rule.

Algorithm 7.3. Choose a sequence $\{\alpha_n\}_{n \in \mathbb{N}} \subset (0, \infty)$ that satisfies
\[
\lim_{n \to \infty} \alpha_n = 0 \quad \text{and} \quad \sum_{n=1}^{\infty} \alpha_n = \infty
\]
Choose an initial feasible point $x_0 \in \mathcal{X}$. For $n = 0, 1, \ldots$ until a chosen stopping criterion is satisfied, do the following steps:
Step 1: Choose an arc $h_n : [0, \infty) \mapsto \mathcal{X}$.

Step 2: Set $x_{n+1} = h_n(\alpha_n)$.

In Algorithm 7.3, $\alpha_n$ may depend on data obtained until iteration $n$. Thus, the sequence \( \{\alpha_n\}_{n \in \mathbb{N}} \) does not have to be predetermined, as long as (189) is satisfied.

As in the case of the constant step-size rule, Algorithm 7.3 describes an algorithm that uses the diminishing step-size rule irrespective of whether $f$ and $\nabla f$ are evaluated exactly or not. However, the choice of arcs $h_n$ change depending on whether $f$ and $\nabla f$ are evaluated exactly or not. If $f$ and $\nabla f$ are evaluated exactly and the scaled gradient projection arc rule in (123) is used, it can be shown that the sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 7.3 converges (e.g., see ?).

In this section, we consider the case when $f$ and $\nabla f$ are approximated by $f_n$ and $\hat{\nabla}_n f$ respectively. We show that the sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 7.3 satisfies the asymptotic descent property. First, we state the conditions that we impose on the arc rule.

A 7.11. There exists $\beta > 0$ such that for each $n$,

\[
\hat{\nabla}_n f(x_n)^T (h_n(\alpha_n) - x_n) \leq -\frac{\beta}{\alpha_n} \|h_n(\alpha_n) - x_n\|^2_2
\]

A 7.12. Consider any sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 7.3, and any subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that $x_{n_k} \to x \in \mathcal{X}$ and $\alpha_{n_k} \to 0$ as $k \to \infty$. Then,

\[
\lim_{k \to \infty} \frac{\|h_{n_k}(\alpha_{n_k}) - x_{n_k}\|}{\alpha_{n_k}} = 0
\]

implies that $x \in \mathcal{S}$.

Next we show that some arc rules satisfy the assumptions stated above.

(a) Consider the scaled gradient projection arc rule defined in (124). That is, at iteration $n$, given $x_n \in \mathcal{X}$ and $H_n \in \mathcal{P}$,

\[
h_n(\alpha) := \Pi_{\mathcal{X}}(x_n - \alpha H_n^{-1} \hat{\nabla}_n f(x_n), H_n)
\]

for each $\alpha \in [0, \infty)$. It follows from (refeqn:proj dir) of Property P 7.5 that for each $\alpha \in (0, \infty)$,

\[
\hat{\nabla}_n f(x_n)^T (h_n(\alpha) - x_n) \leq -\frac{\lambda_{\text{min}}(H_n)}{\alpha} \|h_n(\alpha) - x_n\|^2_2 \leq -\frac{\lambda_1}{\alpha} \|h_n(\alpha) - x_n\|^2_2
\]
and thus Assumption A 7.11 holds with \( \beta = \lambda_1 \). It follows from Lemma 7.6 that Assumption A 7.12 holds.

(b) Consider the scaled gradient inexact projection arc rule defined in (142). That is, at iteration \( n \), given \( x_n \in X \) and \( H_n \in P \),

\[
h_n(\alpha) \in \Pi_X(x_n, \alpha, -H_n^{-1}\hat{\nabla}_n f(x_n), H_n) \tag{190}
\]

for each \( \alpha \in [0, \bar{\alpha}] \). It follows from Corollary 7.9 and (114) that for all \( \alpha \in (0, [(1 - \eta_1)/4]^{1/(\zeta - 2)} \bar{\alpha}] \),

\[
\hat{\nabla}_n f(x_n)^T(h_n(\alpha) - x_n) \leq -\frac{\eta_1 \lambda_1}{\alpha} \|h_n(\alpha) - x_n\|_2^2
\]

and thus Assumption A 7.11 holds with \( \beta = \eta_1 \lambda_1 \). It follows from Lemma 7.15 that Assumption A 7.12 holds.

(c) Consider the feasible direction arc rule defined in (154). That is, at iteration \( n \), given \( x_n, \Phi_n \in X \),

\[
h_n(\alpha) := x_n + \alpha(\Phi_n - x_n)
\]

for each \( \alpha \in [0, 1] \). In Section 7.1.3 we described three target point rules (Examples 7.1 through 7.3) for choosing \( \Phi_n \). First consider the scaled gradient projection target point rule given in (158). That is, at iteration \( n \), given \( \gamma > 0, x_n \in X, \hat{\nabla}_n f(x_n) \in \mathbb{R}^l \), and \( H_n \in \mathbb{R}^{l \times l} \), choose \( \Phi_n := \Pi_X(x_n - \gamma H_n^{-1}\hat{\nabla}_n f(x_n), H_n) \).

It follows from (refeqn:proj dir) of Property P 7.5 that for each \( \gamma \in (0, \infty) \),

\[
\hat{\nabla}_n f(x_n)^T(\Phi_n - x_n) \leq -\frac{\lambda_{\min}(H_n)}{\gamma} \|\Phi_n - x_n\|_2^2 \leq -\frac{\lambda_1}{\gamma} \|\Phi_n - x_n\|_2^2
\]

and thus, for each \( \alpha \in (0, 1] \),

\[
\hat{\nabla}_n f(x_n)^T(h_n(\alpha) - x_n) = \alpha \hat{\nabla}_n f(x_n)^T(\Phi_n - x_n) \leq -\frac{\alpha \lambda_1}{\gamma} \|\Phi_n - x_n\|_2^2 = -\frac{\lambda_1}{\gamma \alpha} \|h_n(\alpha) - x_n\|_2^2
\]

Hence Assumption A 7.11 holds with \( \beta = \lambda_1/\gamma \alpha \). Suppose that \( \lim_{k \to \infty} \|h_{n_k}(\alpha_{n_k}) - x_{n_k}\|_2/\alpha_{n_k} = 0 \). Then

\[
\lim_{k \to \infty} \|\Phi_{n_k} - x_{n_k}\|_2 = 0
\]

\[
\Rightarrow \gamma \lim_{k \to \infty} \frac{\|\Pi_X(x_{n_k} - \gamma H_{n_k}^{-1}\hat{\nabla}_{n_k} f(x_{n_k}), H_{n_k}) - x_{n_k}\|_2}{\gamma} = 0
\]
It follows from the last statement and Lemma 7.6 that Assumption A 7.11 holds.

Next consider the scaled gradient inexact projection target point rule given in (160). That is, at iteration $n$, given $\gamma > 0$, $x_n \in X$, $\nabla_n f(x_n) \in \mathbb{R}^l$, and $H_n \in \mathbb{R}^{l \times l}$, choose

$$
\Phi_n \in \Pi_X(x_n, \gamma, -H_n^{-1}\nabla_n f(x_n), H_n)
$$

It follows from Corollary 7.9 and (114) that for all $\gamma \in (0, [(1 - \eta_1)/4]^{1/(\zeta - 2)} \alpha']$,

$$
\nabla_n f(x_n)^T(\Phi_n - x_n) \leq -\frac{\eta_1 \lambda_1}{\gamma} \|\Phi_n - x_n\|_2^2
$$

and thus, for each $\alpha \in (0, 1]$,

$$
\nabla_n f(x_n)^T(h_n(\alpha) - x_n) = \alpha \nabla_n f(x_n)^T(\Phi_n - x_n) \leq -\frac{\alpha \eta_1 \lambda_1}{\gamma} \|\Phi_n - x_n\|_2^2 = -\frac{\eta_1 \lambda_1}{\gamma \alpha} \|h_n(\alpha) - x_n\|_2^2
$$

Hence Assumption A 7.11 holds with $\beta = \eta_1 \lambda_1/\gamma \alpha$.

Next, we show that under the two assumptions above, the sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.3 satisfies

$$
\lim_{n \to \infty} \sup_{x \in \mathbb{S}} f(x_n) \leq \sup_{x \in \mathbb{S}} f(x)
$$

We first establish a number of useful lemmas.

**Lemma 7.28.** Suppose that $f$ is Lipschitz continuously differentiable with Lipschitz constant $L$ on a neighborhood of $X$. Suppose that the sequence $\{h_n\}_{n \in \mathbb{N}}$ of arcs used in Algorithm 7.3 satisfies Assumption A 7.11. Consider a sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.3. Then, for all $n$,

$$
f(x_n) - f(x_{n+1}) \geq \left( \frac{\beta}{\alpha_n} - \frac{L}{2} \right) \|h_n(\alpha_n) - x_n\|_2 - \|\nabla f(x_n) - \nabla_n f(x_n)\|_2 \|h_n(\alpha_n) - x_n\|_2
$$

**Proof.** Since $f$ is Lipschitz continuously differentiable on $X$ it follows from (182) that

$$
f(x_{n+1}) \leq f(x_n) + \nabla f(x_n)^T(x_{n+1} - x_n) + \frac{L}{2} \|x_{n+1} - x_n\|_2^2
$$

Thus, since $x_{n+1} = h_n(\alpha_n)$,

$$
f(x_n) - f(x_{n+1}) \geq -\nabla_n f(x_n)^T(h_n(\alpha_n) - x_n) - (\nabla f(x_n) - \nabla_n f(x_n))^T(h_n(\alpha_n) - x_n) - \frac{L}{2} \|h_n(\alpha_n) - x_n\|_2^2
$$

It follows from Assumption A 7.11 that there is $\beta > 0$ such that

$$
-\nabla_n f(x_n)^T(h_n(\alpha_n) - x_n) \geq \frac{\beta}{\alpha_n} \|h_n(\alpha_n) - x_n\|_2^2
$$

118
Thus,
\[
f(x_n - f(x_{n+1})) \geq \frac{\beta}{\alpha_n} \|h_n(\alpha_n) - x_n\|^2 - \frac{L}{2} \|h_n(\alpha_n) - x_n\|^2 - \left\langle \nabla f(x_n) - \hat{\nabla} f(x_n), h_n(\alpha_n) - x_n \right\rangle
\]
\[
\geq \left( \frac{\beta}{\alpha_n} - \frac{L}{2} \right) \|h_n(\alpha_n) - x_n\|^2 - \|\nabla f(x_n) - \hat{\nabla} f(x_n)\|_2 \|h_n(\alpha_n) - x_n\|_2
\]
\[
= \left[ \left( \frac{\beta}{\alpha_n} - \frac{L}{2} \right) \|h_n(\alpha_n) - x_n\|_2 - \|\nabla f(x_n) - \hat{\nabla} f(x_n)\|_2 \right] \|h_n(\alpha_n) - x_n\|_2
\]

\[\square\]

Lemma 7.29. Suppose that the function \( f \) satisfies Assumptions A 7.6 and A 7.10, and that the approximating sequences \( \{f_n\}_{n \in \mathbb{N}} \) and \( \{\hat{\nabla} f\}_{n \in \mathbb{N}} \) satisfy Assumption A 5.2. Let \( S \) denote the set of stationary points of \( f \) on \( X \), as defined in (109). Suppose that the sequence \( \{\alpha_n\}_{n \in \mathbb{N}} \) of step sizes satisfies (189) and that the sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs satisfies Assumptions A 7.11 and A 7.12. Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 7.3. Suppose that \( \{x_n\}_{n \in \mathbb{N}} \) is contained in a compact set \( C \subseteq X \) (the set \( C \) can depend on the sequence). Then
\[
\liminf_{n \to \infty} \frac{\|h_n(\alpha_n) - x_n\|_2}{\alpha_n} = 0
\]
In addition, there exists a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subseteq \mathbb{N} \) such that \( x_{n_k} \to x^* \) for some \( x^* \in S \).

Proof. We show the first result by contradiction. Suppose that there is \( \varepsilon > 0 \) and \( N_1 \in \mathbb{N} \) such that for all \( n > N_1 \),
\[
\frac{\|h_n(\alpha_n) - x_n\|_2}{\alpha_n} > \varepsilon
\]
Since \( f \) is continuously differentiable on \( X \), and Assumption A 5.2 holds, and \( \{x_n\}_{n \in \mathbb{N}} \) is contained in a compact subset of \( X \), it follows from Lemma 5.1 that
\[
\lim_{n \to \infty} \left\| \hat{\nabla} f(x_n) - \nabla f(x_n) \right\|_2 = 0
\]
Also, from (189), \( \alpha_n \to 0 \). Thus,
\[
\lim_{n \to \infty} \left( \beta - \frac{L \alpha_n}{2} \right) \varepsilon - \left\| \hat{\nabla} f(x_n) - \nabla f(x_n) \right\|_2 = \beta \varepsilon > 0 \quad (191)
\]
Thus, there exists \( N_2 \in \mathbb{N} \) such that for all \( n > N_2 \),
\[
\left( \beta - \frac{L \alpha_n}{2} \right) \varepsilon - \left\| \hat{\nabla} f(x_n) - \nabla f(x_n) \right\|_2 > \frac{1}{2} \beta \varepsilon > 0
\]
Thus, it follows from Lemma 7.28 that for all \( n > \max\{N_1, N_2\} \),
\[
f(x_n) - f(x_{n+1}) \geq \left[ \left( \frac{\beta}{\alpha_n} - \frac{L}{2} \right) \|h_n(\alpha_n) - x_n\|_2 - \|\nabla f(x_n) - \hat{\nabla} f(x_n)\|_2 \right] \|h_n(\alpha_n) - x_n\|_2 \geq \frac{1}{2} \beta \varepsilon^2 \alpha_n > 0 \quad (192)
\]
Therefore, for all $n > \max\{N_1, N_2\}$, $f(x_n) > f(x_{n+1})$. Thus, $\{f(x_n)\}_{n \in \mathbb{N}}$ is a decreasing sequence for large enough $n$. Hence, since $f$ is bounded below on $\mathcal{X}$, the sequence $\{f(x_n)\}_{n \in \mathbb{N}}$ converges. Thus,

$$\sum_{n=1}^{\infty} (f(x_n) - f(x_{n+1})) < \infty$$

It follows from (192) that

$$\sum_{n=\max\{N_1, N_2\} + 1}^{\infty} \frac{1}{2} \beta^2 \alpha_n \leq \sum_{n=\max\{N_1, N_2\} + 1}^{\infty} (f(x_n) - f(x_{n+1})) < \infty$$

This contradicts the assumption that $\sum_{n=1}^{\infty} \alpha_n = \infty$. Therefore,

$$\liminf_{n \to \infty} \frac{\|h_n(\alpha_n) - x_n\|_2}{\alpha_n} = 0$$

Thus, there is a subsequence $\{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N}$ such that

$$\lim_{k \to \infty} \frac{\|h_{n_k}(\alpha_{n_k}) - x_{n_k}\|_2}{\alpha_{n_k}} = 0$$

Since $\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{C}$ for some compact set $\mathcal{C} \subset \mathcal{X}$, there is a further subsequence $\{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}}$ such that

$$\lim_{k \to \infty} x_{m_k} = x^* \in \mathcal{C} \subset \mathcal{X} \quad \text{and} \quad \lim_{k \to \infty} \frac{\|h_{m_k}(\alpha_{m_k}) - x_{m_k}\|_2}{\alpha_{m_k}} = 0$$

Therefore, it follows from Assumption A 7.12 that $x^* \in \mathcal{S}$. \qed

**Lemma 7.30.** Suppose that the function $f$ satisfies Assumptions A 7.6 and A 7.10, and that the approximating sequences $\{f_n\}_{n \in \mathbb{N}}$ and $\{\nabla_n f\}_{n \in \mathbb{N}}$ satisfy Assumption A 5.2. Suppose that the sequence $\{\alpha_n\}_{n \in \mathbb{N}}$ of step sizes satisfies (189) and that the sequence $\{h_n\}_{n \in \mathbb{N}}$ of arcs satisfies Assumption A 7.11. Consider a sequence $\{x_n\}_{n \in \mathbb{N}}$ generated by Algorithm 7.3. Suppose that $\{x_n\}_{n \in \mathbb{N}}$ is contained in a compact set $\mathcal{C} \subset \mathcal{X}$. Then, for any $\varepsilon > 0$, there exists $N_1(\varepsilon) \in \mathbb{N}$ such that for all $n > N_1(\varepsilon)$,

$$f(x_{n+1}) \leq f(x_n) + \varepsilon$$

**Proof.** It follows from Lemma 7.28 that, for all $n$,

$$f(x_{n+1}) \leq f(x_n) - \left(\frac{\beta}{\alpha_n} - \frac{L}{2}\right) \|h_n(\alpha_n) - x_n\|_2^2 + \left\|\nabla_n f(x_n) - \nabla f(x_n)\right\|_2 \|h_n(\alpha_n) - x_n\|_2$$

Since $\alpha_n \to 0$, there is $N' \in \mathbb{N}$ such that for all $n > N'$, $\alpha_n < 2\beta/L$, and thus $\beta/\alpha_n - L/2 > 0$. Hence, for all $n > N'$,

$$f(x_{n+1}) \leq f(x_n) + \left\|\nabla_n f(x_n) - \nabla f(x_n)\right\|_2 \|h_n(\alpha_n) - x_n\|_2$$

120
Since \( \{x_n\}_{n \in \mathbb{N}} \) is bounded, there exists \( M < \infty \) such that

\[
\|h_n(\alpha_n) - x_n\|_2 = \|x_{n+1} - x_n\|_2 \leq M
\]

for all \( n \). Since \( f \) is continuously differentiable on \( \mathcal{X} \), and Assumption A 5.2 holds, and \( \{x_n\}_{n \in \mathbb{N}} \) is contained in a compact subset of \( \mathcal{X} \), it follows from Lemma 5.1 that

\[
\lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 = 0
\]

Thus, for any \( \varepsilon > 0 \), there is \( N'' \in \mathbb{N} \) such that for all \( n > N'' \),

\[
\left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 \leq \frac{\varepsilon}{M}
\]

Then, for all \( n > N_1(\varepsilon) := \max\{N', N''\} \),

\[
f(x_{n+1}) \leq f(x_n) + \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 \|h_n(\alpha_n) - x_n\|_2 \leq f(x_n) + \varepsilon
\]

\[\square\]

**Lemma 7.31.** Suppose that the function \( f \) satisfies Assumptions A 7.6 and A 7.10, and that the approximating sequences \( \{f_n\}_{n \in \mathbb{N}} \) and \( \{\hat{\nabla}_n f\}_{n \in \mathbb{N}} \) satisfy Assumption A 5.2. Let \( S \) denote the set of stationary points of \( f \) on \( \mathcal{X} \), as defined in (109). Suppose that the sequence \( \{\alpha_n\}_{n \in \mathbb{N}} \) of step sizes satisfies (189) and that the sequence \( \{h_n\}_{n \in \mathbb{N}} \) of arcs satisfies Assumptions A 7.11 and A 7.12. Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 7.3. Then, for any compact set \( \mathcal{D} \subset \mathcal{X} \setminus S \), there exists \( N_2(\mathcal{D}) \in \mathbb{N} \) such that for all \( n > N_2(\mathcal{D}) \), if \( x_n \in \mathcal{D} \), then \( f(x_{n+1}) \leq f(x_n) \).

**Proof.** First, we show by contradiction that there is \( \varepsilon > 0 \) and \( N' \in \mathbb{N} \) such that for all \( n > N' \), if \( x_n \in \mathcal{D} \), then

\[
\|h_n(\alpha_n) - x_n\|_2 > \varepsilon \tag{193}
\]

Suppose that no such \( \varepsilon \) and \( N' \) exist. Then there is a subsequence \( \{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N} \) such that \( x_{n_k} \in \mathcal{D} \) for all \( k \in \mathbb{N} \) and

\[
\lim_{k \to \infty} \|h_{n_k}(\alpha_{n_k}) - x_{n_k}\|_2 = 0
\]

Since \( \mathcal{D} \) is a compact set, there is a further subsequence \( \{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}} \) such that \( x_{m_k} \to x \in \mathcal{D} \) as \( k \to \infty \). Since \( \alpha_n \to 0 \), it follows from Assumption A 7.12 that \( x \in \mathcal{S} \). This contradicts the assumption that \( \mathcal{D} \subset \mathcal{X} \setminus \mathcal{S} \).
Since \( \alpha_n \to 0 \), there is \( N'' \in \mathbb{N} \) such that for all \( n > N'' \), \( \alpha_n < \beta/L \), and thus \( \beta - L\alpha_n/2 > \beta/2 > 0 \). Since \( f \) is continuously differentiable on \( \mathcal{X} \), and Assumption A 5.2 holds, and \( D \subset \mathcal{X} \) is compact, it follows from Lemma 5.1 that there is \( N''' \in \mathbb{N} \) such that for all \( n > N''' \), \( \| \nabla_n f(x_n) - \nabla f(x_n) \|_2 < \beta \varepsilon/2 \).

Thus, it follows from Lemma 7.28 and (193) that for all \( n > \max\{ N', N'', N''' \} \), if \( x_n \in D \), then

\[
\begin{align*}
\| \nabla f(x_n) - \nabla f(x_n) \|_2 &< \beta \varepsilon/2.
\end{align*}
\]

\[\square\]

**Theorem 7.32.** Suppose that the set \( \mathcal{X} \) is compact, that the function \( f \) satisfies Assumptions A 7.6 and A 7.10, and that the approximating sequences \( \{ f_n \}_{n \in \mathbb{N}} \) and \( \{ \nabla_n f \}_{n \in \mathbb{N}} \) satisfy Assumption A 5.2. Let \( S \) denote the set of stationary points of \( f \) on \( \mathcal{X} \), as defined in (109). Suppose that the sequence \( \{ \alpha_n \}_{n \in \mathbb{N}} \) of step sizes satisfies (189), and that the sequence \( \{ h_n \}_{n \in \mathbb{N}} \) of arcs satisfies Assumptions A 7.11 and A 7.12. Then, any sequence \( \{ x_n \}_{n \in \mathbb{N}} \) generated by Algorithm 7.3 satisfies

\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in S} f(x)
\]

In addition, if Assumptions A 6.5 and A 6.6 are satisfied, then

\[
\lim_{n \to \infty} d(x_n, S) = 0
\]

**Proof.** We show the result by verifying that the assumptions of Theorem 6.1 are satisfied.

1. Assumption A 7.6 verifies that Assumption A 6.1 is satisfied.

2. Since \( \mathcal{X} \) is compact, by choosing \( C = \mathcal{X} \), Lemma 7.29 verifies that Assumption A 6.2 is satisfied.

3. Lemma 7.30 verifies that Assumption A 6.3 is satisfied.

4. Let \( \mu := \sup_{x \in S} f(x) < \infty \), and for any \( \varepsilon > 0 \), let \( \mathcal{D}(\varepsilon) := \{ x \in \mathcal{X} : \mu + \varepsilon \leq f(x) \leq \mu + 2\varepsilon \} \).

   It follows from the compactness of \( \mathcal{X} \) and the continuity of \( f \) that \( \mathcal{D}(\varepsilon) \) is compact. It follows from the definitions of \( \mu \) and \( \mathcal{D}(\varepsilon) \) that \( \mathcal{D}(\varepsilon) \subset \mathcal{X} \setminus S \). Then it follows from Lemma 7.31 that there exists \( N_2(\mathcal{D}(\varepsilon)) \in \mathbb{N} \) such that for all \( n > N_2(\mathcal{D}(\varepsilon)) \), if \( x_n \in \mathcal{D}(\varepsilon) \), then \( f(x_{n+1}) \leq f(x_n) \).

   Thus Assumption A 6.4 holds.
Therefore it follows from Theorem 6.1 that

$$\limsup_{n \to \infty} \ f(x_n) \leq \sup_{x \in S} \ f(x)$$

If Assumptions A 6.5 and A 6.6 are satisfied, then it follows from Corollary 6.4 that

$$\lim_{n \to \infty} \ d(x_n, S) = 0$$

8 Trust Region Algorithms

In this section, we consider trust region algorithms with inexact function and derivative estimates. A typical trust region algorithm used for a minimization problem \((P)\) works as follows when the function values and derivatives are computed exactly.

**Algorithm 8.1.** Choose constants

$$0 < \eta_1 \leq \eta_2 < 1, \ 0 < \sigma_1 \leq \sigma_2 < 1, \ \Delta_{\text{max}} > 0$$

Choose initial trust region radius \(\Delta_0 \in (0, \Delta_{\text{max}})\). Choose an initial feasible point \(x_0 \in \mathcal{X}\). For \(n = 0, 1, \ldots\) until a chosen stopping criterion is satisfied, do the following steps:

**Step 1:** Choose a model function \(m_n : \mathcal{X} \to \mathbb{R}\) and a trust region \(T_n := \{x_n + d : \|d\|_n \leq \Delta_n\}\).

**Step 2:** Find \(x_n + d_n \in \mathcal{X} \cap T_n\) such that \(m_n(x_n) - m_n(x_n + d_n)\) is greater than a specified positive value.

**Step 3:** Evaluate

$$\rho_n := \frac{f(x_n) - f(x_n + d_n)}{m_n(x_n) - m_n(x_n + d_n)}$$

If \(\rho_n \geq \eta_1\), then set \(x_{n+1} = x_n + d_n\); else set \(x_{n+1} = x_n\).

**Step 4:** Update the trust region radius as follows:

$$\Delta_{n+1} \in \begin{cases} 
[\Delta_n, \Delta_{\text{max}}] & \text{if } \rho_n \geq \eta_2 \\
[\sigma_2 \Delta_n, \Delta_n] & \text{if } \rho_n \in (\eta_1, \eta_2) \\
[\sigma_1 \Delta_n, \sigma_2 \Delta_n] & \text{if } \rho_n < \eta_1 
\end{cases}$$
In the algorithm above, $\Delta_n$ is called the **trust region radius**. Also, $\|\cdot\|_n$ is called the **trust region norm**, and the subscript $n$ in $\|\cdot\|_n$ indicates that the chosen trust region norm may depend on the data collected up to iteration $n$, as long as all the norms used are uniformly equivalent. Details on how the trust region norm can be changed at each iteration are given in [?]. In this section, we use the same trust region norm in all iterations. All norms on $\mathbb{R}^l$ are equivalent, that is, for each norm $\|\cdot\|$ there exists a constant $C \geq 1$ such that for all $x \in \mathbb{R}^l$,

$$\frac{1}{C} \|x\| \leq \|x\|_2 \leq C \|x\|$$

(194)

where $\|\cdot\|_2$ denotes the usual Euclidean norm. Nevertheless, the choice of trust region norm is important for computational purposes because the trust region norm used for a particular problem can have a great impact on the performance of the algorithm.

**Vijay**: I think there is an equivalence between the choice of $H_n$ below and the choice of $\|\cdot\|$ if, as is usually the case, $\|\cdot\|$ is given by $\|x\| = \sqrt{x^T Q x}$ for some $Q$.

Next, we describe model functions $m_n : \mathcal{X} \mapsto \mathbb{R}$ that approximate the function $f$ in $\mathcal{X} \cap T_n$. When $f$ and $\nabla f$ are evaluated exactly, $m_n$ is often defined to be a quadratic function, as follows:

$$m_n(x) := f(x_n) + \nabla f(x_n)^T (x - x_n) + \frac{1}{2} (x - x_n)^T H_n (x - x_n)$$

(195)

where $H_n \in \mathbb{S}^{l \times l}$. Further, if $\nabla^2 f(x_n)$ is also evaluated exactly, then one can choose $H_n = \nabla^2 f(x_n)$.

After the model function and trust region have been chosen, an approximate minimizer $x_n + d_n$ of $m_n$ on $\mathcal{X} \cap T_n$ is found, that satisfies a minimum improvement condition that we will elaborate on later. The relative improvement $\rho_n$ is evaluated, and if it is sufficient, then the iterate is updated by setting $x_{n+1} = x_n + d_n$. Also, depending on the relative improvement, the trust region radius $\Delta_n$ is either reduced, left unchanged, or increased for the next iteration.

Next, suppose that neither $f(x)$ nor its derivatives are evaluated exactly. Instead, we generate sequences $\{f_n\}_{n \in \mathbb{N}}, \{\nabla_n f\}_{n \in \mathbb{N}},$ and $\{\nabla^2_n f\}_{n \in \mathbb{N}}$ that approximate respectively $f$, $\nabla f$ and possibly $\nabla^2 f$. We will design convergent trust region algorithms that use these approximations. Before we describe our modified trust region algorithm, we first state our assumptions regarding the optimization problem and the approximating sequences.

We assume that the feasible set $\mathcal{X}$ satisfies Assumption A 7.1 and the objective function $f$ satisfies Assumptions A 5.1 and A 7.10. Further, we also require some conditions regarding the type of convergence of the approximating sequences.
The sequence \( \{f_n\}_{n \in \mathbb{N}} \subset \mathcal{W}_0(X) \), and for every compact set \( D \subset X \),

\[
\lim_{n \to \infty} \| f_n - f \|_{\mathcal{W}_0(D)} = 0
\]

Recall that the convergence in Lipschitz norm used above means that

\[
\lim_{n \to \infty} \sup_{x \in D} |f_n(x) - f(x)| = 0 \quad \text{and} \quad \lim_{n \to \infty} \sup_{\{x,y \in D : x \neq y\}} \left| \frac{[f_n(y) - f(y)] - [f_n(x) - f(x)]}{\|y-x\|_2} \right| = 0
\]

For any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset X \) such that \( x_n \to x \in X \) as \( n \to \infty \), it holds that

\[
\lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(x) \right\|_2 = 0
\]

Note that it follows from Lemma 5.1 that for any sequence \( \{x_n\}_{n \in \mathbb{N}} \) contained in a compact subset of \( X \),

\[
\lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 = 0
\]

For the convergence analysis of our algorithm, we do not assume any particular form for the model function \( m_n \) that the algorithm uses at iteration \( n \). We will only require that each \( m_n \) satisfies the following two assumptions.

The function \( m : X \to \mathbb{R} \) is twice continuously differentiable on a neighborhood of the feasible set \( X \). Further, \( m_n(x_n) = f_n(x_n) \) and \( \nabla m_n(x_n) = \hat{\nabla}_n f(x_n) \).

For each \( n \in \mathbb{N} \), let \( \kappa_n \) be given by

\[
\kappa_n := 1 + \sup_{x \in X} \left\| \nabla^2 m_n(x) \right\|_2
\]

Then the sequence \( \{\kappa_n\}_{n \in \mathbb{N}} \) is bounded, i.e., there exists \( \kappa_{\max} < \infty \) such that \( \kappa_n \leq \kappa_{\max} \) for all \( n \in \mathbb{N} \).

At each iteration \( n \), the following model function \( m_n \) can be used:

\[
m_n(x) := f_n(x_n) + \hat{\nabla}_n f(x_n)^T (x - x_n) + \frac{1}{2} (x - x_n)^T \hat{\nabla}_n^2 f(x_n) (x - x_n)
\]

It is easy to check that Assumption A 8.3 is satisfied when \( m_n \) is chosen as in (197). Also, if \( m_n \) is chosen as in (197), then Assumption A 8.4 is equivalent to the assumption that there exists \( K_H < \infty \) such that for all \( n \in \mathbb{N} \), \( \sup_{x \in X} \left\| \hat{\nabla}_n^2 f(x) \right\|_2 \leq K_H \).
In Step 3 of Algorithm 8.1, the calculation of $\rho_n$ requires the function evaluations $f(x_n)$ and $f(x_n + d_n)$. Since in our case of interest, $f$ is not evaluated exactly, we use $f_n$ instead and define

$$
\rho_n := \frac{f_n(x_n) - f_n(x_n + d_n)}{m_n(x_n) - m_n(x_n + d_n)}
$$

Our definition of the trust region is the same as that in Algorithm 8.1, i.e.,

$$
T_n := \{ x_n + d \in \mathcal{X} : \|d\| \leq \Delta_n \} \quad (198)
$$

Recall that we wish to design algorithms that search for points in the set $\mathcal{S}$ of stationary points of $f$ on $\mathcal{X}$. We have already seen the following equivalent representations of $\mathcal{S}$:

$$
\mathcal{S} = \{ x \in \mathcal{X} : \nabla f(x)^T (y - x) \geq 0 \ \forall \ y \in \mathcal{X} \}
$$

$$
= \{ x \in \mathcal{X} : x = \Pi_{\mathcal{X}}(x - \alpha H^{-1}\nabla f(x), H) \ \forall \ \alpha \geq 0 \ \text{and} \ H \in S_{++}^{l \times l} \}
$$

Such characterizations of the set of stationary points can be generalized using the concept of optimality functions, which we discuss next.

**Definition 8.1.** Let the set $\mathcal{X} \subset \mathbb{R}^l$ be closed and convex. A function $\Gamma_{\mathcal{X}} : \mathcal{X} \times \mathbb{R}^l \mapsto [0, \infty)$ is called an **optimality function** for the set $\mathcal{X}$ if it satisfies the following two properties.

**P 8.1.** For any function $g : \mathcal{X} \mapsto \mathbb{R}$ that is continuously differentiable on a neighborhood of $\mathcal{X}$ and any $x \in \mathcal{X}$, it holds that $\Gamma_{\mathcal{X}}(x, \nabla g(x)) = 0$ if and only if $x$ is a stationary point of $g$ in $\mathcal{X}$; i.e.,

$$
\Gamma_{\mathcal{X}}(x, \nabla g(x)) = 0 \iff \nabla g(x)^T (y - x) \geq 0 \ \forall \ y \in \mathcal{X}
$$

**P 8.2.** The function $\Gamma_{\mathcal{X}}$ is lower semi-continuous in both its arguments, i.e., for any sequence $\{(x_n, d_n)\}_{n \in \mathbb{N}} \subset \mathcal{X} \times \mathbb{R}^l$ such that $(x_n, d_n) \to (x, d)$ as $n \to \infty$, it holds that

$$
\liminf_{n \to \infty} \Gamma_{\mathcal{X}}(x_n, d_n) \geq \Gamma_{\mathcal{X}}(x, d)
$$

An example of an optimality function $\Gamma_{\mathcal{X}} : \mathcal{X} \times \mathbb{R}^l \mapsto [0, \infty)$ that satisfies the properties above is $\Gamma_{\mathcal{X}}(x, d) := \|\Pi_{\mathcal{X}}(x - \alpha H^{-1}d, H) - x\|_2$ for some $\alpha > 0$ and $H \in S_{++}^{l \times l}$. Given any optimality function $\Gamma_{\mathcal{X}}$ for $\mathcal{X}$, since we have assumed that $f$ is continuously differentiable on $\mathcal{X}$, we can use Property P 8.1 and define $\mathcal{S}$ as

$$
\mathcal{S} := \{ x \in \mathcal{X} : \Gamma_{\mathcal{X}}(x, \nabla f(x)) = 0 \}
$$

(199)

The following lemma is an immediate consequence of the above two properties of $\Gamma_{\mathcal{X}}$. 

126
Lemma 8.1. Suppose that \( \{(x_n,d_n)\}_{n \in \mathbb{N}} \subset \mathcal{X} \times \mathbb{R}^l \) satisfies \( (x_n,d_n) \to (x,\nabla f(x)) \) as \( n \to \infty \). If

\[
\liminf_{n \to \infty} \Gamma_x(x_n,d_n) = 0
\]

then \( x \in \mathcal{S} \).

Proof. It follows from the lower semi-continuity of \( \Gamma_x \) that

\[
\Gamma_x(x,\nabla f(x)) \leq \liminf_{n \to \infty} \Gamma_x(x_n,d_n) = 0
\]

Since \( \Gamma_x(x,\nabla f(x)) \geq 0 \) for all \( x \in \mathcal{X} \), it follows that \( \Gamma_x(x,\nabla f(x)) = 0 \). Thus \( x \in \mathcal{S} \).

Consider any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that \( \Gamma_x(x_n,\nabla f(x_n)) \to 0 \) as \( n \to \infty \). It follows from Lemma 8.1 that all the limit points of \( \{x_n\}_{n \in \mathbb{N}} \) are contained in \( \mathcal{S} \). Next we develop an algorithm that generates sequences \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) such that \( \Gamma_x(x_n,\nabla f(x_n)) \to 0 \) as \( n \to \infty \).

Algorithm 8.2. Choose constants

\[
0 < \eta_1 \leq \eta_2 < 1, \quad 0 < \sigma_1 \leq \sigma_2 < 1 < \sigma_3, \quad \Delta_{\text{max}} > 0
\]

\[
0 < \mu_1 \leq 1
\]

Choose an optimality function \( \Gamma_x \) for \( \mathcal{X} \) that satisfies Properties P 8.1 and P 8.2. Choose initial trust region radius \( \Delta_0 \in (0,\Delta_{\text{max}}) \). Choose an initial feasible point \( x_0 \in \mathcal{X} \). For \( n = 0,1,\ldots \) until a chosen stopping criterion is satisfied, do the following steps:

**Step 1:** Choose a model function \( m_n : \mathcal{X} \to \mathbb{R} \) that satisfies Assumptions A 8.3 and A 8.4 and a trust region \( T_n := \{x_n + d : \|d\| \leq \Delta_n\} \).

**Step 2:** If \( x_n \) is a stationary point of \( m_n \) in \( \mathcal{X} \cap T_n \), then set \( x_{n+1} = x_n, \Delta_{n+1} = \Delta_n \) and return to Step 1.

**Step 3:** Find a point \( x_n + d_n \in \mathcal{X} \cap T_n \) such that

\[
m_n(x_n) - m_n(x_n + d_n) \geq \mu_1 \Gamma_x(x_n,\nabla m_n(x_n)) \min \left\{1, \Delta_n, \frac{\Gamma_x(x_n,\nabla m_n(x_n))}{\kappa_{\text{max}}} \right\}
\]

(200)

**Step 4:** Evaluate

\[
\rho_n := \frac{f_n(x_n) - f_n(x_n + d_n)}{m_n(x_n) - m_n(x_n + d_n)}
\]

If \( \rho_n \geq \eta_1 \), then set \( x_{n+1} = x_n + d_n \); else set \( x_{n+1} = x_n \).
Step 5: Update the trust region radius as follows:

\[
\Delta_{n+1} = \begin{cases} 
\Delta_n, \min\{\sigma_3 \Delta_n, \Delta_{\text{max}}\} & \text{if } \rho_n \geq \eta_2 \\
[\sigma_2 \Delta_n, \Delta_n] & \text{if } \rho_n \in [\eta_1, \eta_2) \\
[\sigma_1 \Delta_n, \sigma_2 \Delta_n] & \text{if } \rho_n < \eta_1 
\end{cases}
\]

Steps 2 and 3 in Algorithm 8.2 warrant the following comments. In Step 2, we did not mention how we can determine whether \(x_n\) is a stationary point of \(m_n\) in \(\mathcal{X} \cap T_n\) or not. The following lemma shows that we can use the function \(\Gamma_x\) to do so.

**Lemma 8.2.** Suppose that the set \(\mathcal{X}\) is nonempty, closed, and convex, and that the model function \(m_n\) satisfies Assumption A 8.3. Then \(x_n \in \mathcal{X} \cap T_n\) is a stationary point of \(m_n\) in \(\mathcal{X}\) if and only if \(x_n\) is a stationary point of \(m_n\) in \(\mathcal{X} \cap T_n\).

**Proof.** First, note that \(\mathcal{X} \cap T_n\) is closed and convex. Also, since \(m_n\) satisfies Assumption A 8.3, it is continuously differentiable on \(\mathcal{X} \cap T_n\).

Suppose that \(x_n\) is a stationary point of \(m_n\) in \(\mathcal{X}\). Then \(\nabla m_n(x_n)^T(x - x_n) \geq 0\) for all \(x \in \mathcal{X}\). In particular, \(\nabla m_n(x_n)^T(x - x_n) \geq 0\) for all \(x \in \mathcal{X} \cap T_n\). Hence, \(x_n\) is a stationary point of \(m_n\) in \(\mathcal{X} \cap T_n\).

Suppose that \(x_n\) is a stationary point of \(m_n\) in \(\mathcal{X} \cap T_n\). Then \(\nabla m_n(x_n)^T(x - x_n) \geq 0\) for all \(x \in \mathcal{X} \cap T_n\). Since \(\Delta_n > 0\), the interior of \(T_n\) is nonempty. Thus, since \(x_n\) lies at the center of the trust region \(T_n := \{x \in \mathbb{R}^l : \|x - x_n\| \leq \Delta_n\}\), it is clear that \(x_n\) is in the interior of \(T_n\). Hence, for any \(x \in \mathbb{R}^l\), there is \(\varepsilon_x \in (0,1)\) small enough such that \(x_n + \varepsilon_x(x - x_n) \in T_n\). In particular, consider any \(x \in \mathcal{X}\). Since \(\mathcal{X}\) is convex and \(x_n \in \mathcal{X}\), \(x_n + \varepsilon_x(x - x_n) \in \mathcal{X}\). Thus, for any \(x \in \mathcal{X}\), there is \(\varepsilon_x > 0\) small enough such that \(x_n + \varepsilon_x(x - x_n) \in \mathcal{X} \cap T_n\). Hence, since \(x_n\) is a stationary point of \(m_n\) in \(\mathcal{X} \cap T_n\),

\[
\nabla m_n(x_n)^T(x - x_n) = \frac{1}{\varepsilon_x} \nabla m_n(x_n)^T([x_n + \varepsilon_x(x - x_n)] - x_n) \geq 0
\]

Thus, \(\nabla m_n(x_n)^T(x - x_n) \geq 0\) for all \(x \in \mathcal{X}\), and therefore \(x_n\) is a stationary point of \(m_n\) in \(\mathcal{X}\). \(\Box\)

Thus, one can determine whether \(x_n\) is a stationary point of \(m_n\) in \(\mathcal{X} \cap T_n\) by evaluating \(\Gamma_x(x_n, \nabla m_n(x_n))\), because it follows from Property P 8.1 and Lemma 8.2 that \(x_n\) is a stationary point of \(m_n\) in \(\mathcal{X} \cap T_n\) if and only if \(\Gamma_x(x_n, \nabla m_n(x_n)) = 0\).
Note that if \( x_n \) is not a stationary point of \( m_n \) in \( \mathcal{X} \cap \mathcal{T}_n \), then Step 3 requires us to find a point \( x_n + d_n \in \mathcal{X} \cap \mathcal{T}_n \) such that (200) is satisfied. We have however, not provided a particular method to find such a point. We do so in order to keep our analysis as general as possible, i.e., so that we may be able to include a wide variety of algorithms into the framework of Algorithm 8.2. We will show later in our convergence analysis that whereas a particular implementation of Algorithm 8.2 may use any method to find \( x_n + d_n \), as long as \( x_n + d_n \) satisfies (200) in each iteration, the algorithm is guaranteed to converge to a stationary point of \( f \) in \( \mathcal{X} \). Admittedly, it is critical for the progress of Algorithm 8.2, that whenever \( x_n \) is not a stationary point of \( m_n \) in \( \mathcal{X} \cap \mathcal{T}_n \), there must exist at least one point \( x_n + d_n \in \mathcal{X} \cap \mathcal{T}_n \) that satisfies (200). In this paper, we do not study conditions on \( m_n \) and \( \Gamma_X \) that guarantee the existence of such a point. Instead we make the following assumption on its existence and thereafter provide a two examples of \( \Gamma_X \) and \( m_n \) for which such an assumption is valid.

A 8.5. For each point \( x_n \) that is not a stationary point of \( m_n \) in \( \mathcal{X} \cap \mathcal{T}_n \), there exists a point \( x_n^\ast \in \mathcal{X} \cap \mathcal{T}_n \) that satisfies
\[
m_n(x_n) - m_n(x_n^\ast) \geq \mu_1 \Gamma_X(x_n, \nabla m_n(x_n)) \min \left\{ 1, \Delta_n, \frac{\Gamma_X(x_n, \nabla m_n(x_n))}{\kappa_{\max}} \right\}
\]
(201)

Next we consider two widely-used examples of optimality functions \( \Gamma_X \) and indicate how a point \( x_n^\ast \) satisfying (201) can be found for any non-stationary \( x_n \in \mathcal{X} \) for each \( \Gamma_X \).

Example 8.1. Suppose that the set \( \mathcal{X} \) is nonempty, compact and convex. Consider the optimality function \( \Gamma_X : \mathcal{X} \times \mathbb{R}^l \mapsto \mathbb{R} \) given by
\[
\Gamma_X(x, d) := -\min_{y \in \mathcal{X}} d^T(y - x)
\]
(202)

The following result shows that \( \Gamma_X \) defined in (202) satisfies Properties P 8.1 and P 8.2.

Proposition 8.3. Suppose that the set \( \mathcal{X} \) is nonempty, compact and convex. Then the function \( \Gamma_X \) defined in (202) satisfies Properties P 8.1 and P 8.2.

Proof. First we show that \( \Gamma_X \) is satisfies Property P 8.1. Note that for any \( x \in \mathcal{X} \) and \( d \in \mathbb{R}^l \),
\[
\min_{y \in \mathcal{X}} d^T(y - x) \leq 0
\]
because \( d^T(x - x) = 0 \). Hence \( \Gamma_X(x, d) \geq 0 \) for all \( x \in \mathcal{X} \) and \( d \in \mathbb{R}^l \). Consider any function \( g \) that is continuously differentiable on \( \mathcal{X} \). It follows immediately from the definition of \( \Gamma_X \) that \( \Gamma_X(x, \nabla g(x)) = 0 \) if and only if \( \nabla g(x)^T(y - x) \geq 0 \) for all \( y \in \mathcal{X} \). Thus, \( \Gamma_X \) satisfies Property P 8.1.
Next, we show that $\Gamma_{x}$ satisfies Property P 8.2. Consider any sequence $\{(x_n,d_n)\}_{n\in\mathbb{N}} \subset \mathcal{X} \times \mathbb{R}^l$ such that $(x_n,d_n) \to (x,d)$ as $n \to \infty$. Note that for any $x' \in \mathcal{X}$,

$$d^T_n(x' - x_n) \geq \min_{y \in \mathcal{X}} \ d^T_n(y - x_n) = -\Gamma_{x}(x_n,d_n)$$

It follows from taking the limit as $n \to \infty$ on both sides that

$$d^T(x' - x) \geq -\liminf_{n \to \infty} \Gamma_{x}(x_n,d_n) \quad \forall \ x' \in \mathcal{X}$$

$$\Rightarrow \ \min_{x' \in \mathcal{X}} d^T(x' - x) \geq -\liminf_{n \to \infty} \Gamma_{x}(x_n,d_n)$$

This shows that $\Gamma_{x}(x,d) := -\min_{x' \in \mathcal{X}} d^T(x' - x) \leq \liminf_{n \to \infty} \Gamma_{x}(x_n,d_n)$. \hfill \Box

Next we discuss how, given $x_n \in \mathcal{X} \cap \mathcal{T}_n$ that is not a stationary point of $m_n$ in $\mathcal{X}$, one can find a point $x'_n \in \mathcal{X} \cap \mathcal{T}_n$ that satisfies (201) with respect to $\Gamma_{x}$ defined above. Given $m_n$ and $x_n \in \mathcal{X} \cap \mathcal{T}_n$, find a target point $\Phi_n \in \mathcal{X}$ that satisfies

$$\Phi_n \in \arg\min_{y \in \mathcal{X}} \nabla m_n(x_n)^T(y - x_n) \quad (203)$$

Since $\mathcal{X}$ is compact and $\nabla m_n(x_n)^T(y - x_n)$ is continuous in $y$, the right side of (203) is nonempty for all $x_n$. Also, it follows from Assumption A 8.3 that $\nabla m_n(x_n) = \nabla f_n(x_n)$. Then,

$$\Gamma_{x}(x_n,\nabla m_n(x_n)) := -\min_{y \in \mathcal{X}} \nabla m_n(x_n)^T(y - x_n) = -\nabla m_n(x_n)^T(\Phi_n - x_n) \quad (204)$$

Next, define the corresponding arc $h_n : [0,1] \mapsto \mathcal{X} \cap \mathcal{T}_n$ by

$$h_n(\alpha) := \begin{cases} x_n + \alpha \ \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} (\Phi_n - x_n) \quad \text{if} \ \Phi_n \neq x_n \\ x_n \quad \text{otherwise} \end{cases} \quad (205)$$

If $\Phi_n \neq x_n$, then for all $\alpha \in [0,1]$, it follows from the convexity of $\mathcal{X}$ that $h_n(\alpha) \in \mathcal{X}$, and

$$\|h_n(\alpha) - x_n\| = \alpha \ \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \|\Phi_n - x_n\| \leq \min \{\Delta_n, \|\Phi_n - x_n\|\} \leq \Delta_n$$

If $\Phi_n = x_n$, then $h_n(\alpha) = x_n$ for all $\alpha$. Thus, for all $\alpha \in [0,1]$, $h_n(\alpha) \in \mathcal{X} \cap \mathcal{T}_n$.

Next, we find an Armijo step-size $\alpha_n^\tau \in [0,1]$ along the arc $h_n$ that satisfies the Armijo sufficient conditions (163)–(165), as applied to the function $m_n$. In particular, for some $\eta, \tau \in (0,1)$, we find $\alpha_n^\tau$ that satisfies

$$m_n(x_n) - m_n(h_n(\alpha_n^\tau)) \geq -\eta \nabla m_n(x_n)^T(h_n(\alpha_n^\tau) - x_n) \quad (206)$$
and
\[ \alpha^*_n \in [\tau \bar{\alpha}_n, 1] \]  \hspace{1cm} (207)
where \( \bar{\alpha}_n \) satisfies either
\[ m_n(x_n) - m_n(h_n(\bar{\alpha}_n)) < -\eta \nabla m_n(x_n)^T(h_n(\bar{\alpha}_n) - x_n) \]  \hspace{1cm} (208)
or \( \bar{\alpha}_n \geq 1 \). The following lemma establishes that there exists an Armijo step-size \( \alpha^*_n \) that satisfies (206)–(208) whenever \( x_n \) is not a stationary point of \( m_n \) in \( X \).

**Lemma 8.4.** Suppose that the set \( X \) is nonempty, compact, and convex, and that the model function \( m_n : X \mapsto \mathbb{R} \) satisfies Assumption A 8.3. Consider any \( x_n \in X \) that is not a stationary point of \( m_n \) in \( X \). Consider any target point \( \Phi_n \in X \) that satisfies (203). Let arc \( h_n : [0, 1] \mapsto X \cap T_n \) be given by (205). Then there exists a step-size \( \alpha^*_n \in [0, 1] \) that satisfies (206)–(208).

**Proof.** We establish the result by showing that all the conditions of Lemma 7.22 are satisfied. First, it follows from Assumption A 8.3 that \( m_n \) is continuously differentiable. Since \( x_n \) is not a stationary point of \( m_n \) in \( X \), \( \Phi_n \neq x_n \), and hence
\[ h_n(\alpha) = x_n + \alpha \min \left\{ \frac{\Delta_n}{\| \Phi_n - x_n \|}, 1 \right\} (\Phi_n - x_n) \quad \text{for all} \quad \alpha \in [0, 1] \]
Next we show that \( h_n \) satisfies (166)–(168).
\[ \lim_{\alpha \downarrow 0} \| h_n(\alpha) - x_n \|_2 = \min \left\{ \frac{\Delta_n}{\| \Phi_n - x_n \|}, 1 \right\} \| \Phi_n - x_n \|_2 \lim_{\alpha \downarrow 0} \alpha = 0 \]
Thus, \( h_n \) satisfies (166).

Since \( x_n \) is not a stationary point of \( m_n \) in \( X \), \( \nabla m_n(x_n)^T(\Phi_n - x_n) < 0 \). Hence
\[ \nabla m_n(x_n)^T(h_n(\alpha) - x_n) = \alpha \min \left\{ \frac{\Delta_n}{\| \Phi_n - x_n \|}, 1 \right\} \nabla m_n(x_n)^T(\Phi_n - x_n) < 0 \]
for all \( \alpha \in (0, 1] \). Thus, \( h_n \) satisfies (167).

Also,
\[ \limsup_{\alpha \downarrow 0} \nabla m_n(x_n)^T \frac{h_n(\alpha) - x_n}{\| h_n(\alpha) - x_n \|_2} = \nabla m_n(x_n)^T \frac{\Phi_n - x_n}{\| \Phi_n - x_n \|_2} < 0 \]
Therefore, \( h_n \) satisfies (168). Hence, it follows from Lemma 7.22 that there exists \( \alpha^*_n \in [0, 1] \) that satisfies (206)–(208). \qed
Next we show that \( x_n^\Gamma := h_n(\alpha_n^\Gamma) \), where \( \alpha_n^\Gamma \) satisfies the Armijo sufficient conditions, satisfies (201) for the optimality function \( \Gamma_X \).

**Lemma 8.5.** Suppose that the set \( \mathcal{X} \) is nonempty, compact, and convex and that the model function \( m_n : \mathcal{X} \mapsto \mathbb{R} \) satisfies Assumptions A 8.3 and A 8.4. Consider any \( x_n \in \mathcal{X} \) that is not a stationary point of \( m_n \) in \( \mathcal{X} \). Consider any target point \( \Phi_n \in \mathcal{X} \) that satisfies (203). Let arc \( h_n : [0, 1] \mapsto \mathcal{X} \cap T_n \) be given by (205), and let \( \alpha_n^r \) be any Armijo step-size that satisfies (206)–(208). Then \( x_n^\Gamma := h_n(\alpha_n^r) \) satisfies (201) with

\[
0 < \mu_1 \leq \min \left\{ \eta \tau, \frac{\eta \tau (1 - \eta) \tau}{(BC)^2} \right\}
\]

where \( \eta, \tau \in (0, 1) \) are the parameters associated with the Armijo step-size \( \alpha_n^r \), \( C \) is the constant in (194), and \( B := \sup_{x, y \in \mathcal{X}} \| y - x \| > 0 \) is the diameter of \( \mathcal{X} \).

**Proof.** Since \( x_n \) is not a stationary point of \( m_n \) in \( \mathcal{X} \), it follows from the definition of \( h_n \) in (205), and from the definition of \( \Gamma_X \) in (204) that

\[
\nabla m_n(x_n)^T (h_n(\alpha_n^r) - x_n) = \alpha_n^r \min \left\{ \frac{\Delta_n}{\| \Phi_n - x_n \|}, 1 \right\} \nabla m_n(x_n)^T (\Phi_n - x_n) = -\alpha_n^r \min \left\{ \frac{\Delta_n}{\| \Phi_n - x_n \|}, 1 \right\} \Gamma_X(x_n, \nabla m_n(x_n))
\]

Consider the following cases regarding \( \alpha_n^r \).

**Case(a)** Suppose that \( \alpha_n^r \in [\tau \bar{\alpha}_n, 1] \) and \( \bar{\alpha}_n \geq 1 \). Then \( \alpha_n^r \geq \tau \), and it follows from (206) that

\[
m_n(x_n) - m_n(h_n(\alpha_n^r)) \geq \eta \tau \min \left\{ \frac{\Delta_n}{\| \Phi_n - x_n \|}, 1 \right\} \Gamma_X(x_n, \nabla m_n(x_n)) \geq \eta \tau \min \left\{ \frac{\Delta_n}{B}, 1 \right\} \Gamma_X(x_n, \nabla m_n(x_n)) \tag{209}
\]

**Case(b)** Suppose that \( \alpha_n^r \in [\tau \bar{\alpha}_n, 1] \), and \( \bar{\alpha}_n < 1 \) satisfies (208). For brevity of notation, let

\[
\hat{\alpha}_n := \bar{\alpha}_n \min \left\{ \frac{\Delta_n}{\| \Phi_n - x_n \|}, 1 \right\}
\]

Since \( m_n \) is twice continuously differentiable on \( \mathcal{X} \), there is \( y = x_n + t \hat{\alpha}_n (\Phi_n - x_n) \) for some \( t \in (0, 1) \) such that

\[
m_n(x_n + \hat{\alpha}_n (\Phi_n - x_n)) = m_n(x_n) + \hat{\alpha}_n \nabla m_n(x_n)^T (\Phi_n - x_n) + \frac{1}{2} \hat{\alpha}_n^2 (\Phi_n - x_n)^T \nabla^2 m_n(y) (\Phi_n - x_n)
\]

132
Note that $x_n + \hat{\alpha}_n (\Phi_n - x_n) = h_n (\hat{\alpha}_n) \in \mathcal{X} \cap \mathcal{T}_n$. Since $x_n \in \mathcal{X} \cap \mathcal{T}_n$ and $\mathcal{X} \cap \mathcal{T}_n$ is convex, it follows that $y \in \mathcal{X} \cap \mathcal{T}_n$. It follows from Assumption A 8.4 that

$$m_n(h_n(\hat{\alpha}_n)) - m_n(x_n) - \hat{\alpha}_n \nabla m_n(x_n)^T(\Phi_n - x_n) \leq \frac{1}{2} \hat{\alpha}_n^2 \kappa_{\max} \|\Phi_n - x_n\|^2 \tag{210}$$

Since $\hat{\alpha}_n$ satisfies (208), it also holds that

$$m_n(h_n(\hat{\alpha}_n)) - m_n(x_n) > \eta \hat{\alpha}_n \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \nabla m_n(x_n)^T(\Phi_n - x_n)$$

$$= \eta \hat{\alpha}_n \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \nabla m_n(x_n)^T(\Phi_n - x_n) \tag{211}$$

Recall that $-\nabla m_n(x_n)^T(\Phi_n - x_n) = \Gamma_x(x_n, \nabla m_n(x_n))$. It follows from (210), (211), and (194) that

$$\frac{1}{2} C^2 \hat{\alpha}_n^2 \kappa_{\max} B^2 \geq \frac{1}{2} C^2 \hat{\alpha}_n^2 \kappa_{\max} \|\Phi_n - x_n\|^2$$

$$\geq \frac{1}{2} \hat{\alpha}_n \kappa_{\max} \|\Phi_n - x_n\|^2$$

$$\geq -(1 - \eta) \hat{\alpha}_n \nabla m_n(x_n)^T(\Phi_n - x_n)$$

$$= (1 - \eta) \hat{\alpha}_n \Gamma_x(x_n, \nabla m_n(x_n))$$

$$\Rightarrow \hat{\alpha}_n > \frac{2(1 - \eta)}{B^2 C^2 \kappa_{\max}} \Gamma_x(x_n, \nabla m_n(x_n))$$

It follows from the Armijo condition (206) that

$$m_n(x_n) - m_n(h_n(\alpha^r_n)) \geq -\eta \nabla m_n(x_n)^T(h_n(\alpha^r_n) - x_n)$$

$$= \eta \alpha^r_n \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \Gamma_x(x_n, \nabla m_n(x_n))$$

$$\geq \eta \tau \hat{\alpha}_n \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \Gamma_x(x_n, \nabla m_n(x_n))$$

$$= \eta \tau \hat{\alpha}_n \Gamma_x(x_n, \nabla m_n(x_n))$$

$$> \frac{2\eta \tau (1 - \eta)}{B^2 C^2 \kappa_{\max}} \Gamma_x(x_n, \nabla m_n(x_n))^2 \tag{212}$$

Combining (209) and (212), it follows that

$$m_n(x_n) - m_n(h_n(\alpha^r_n)) \geq \min \left\{ \eta \tau \min \left\{ \frac{\Delta_n}{B}, 1 \right\} \Gamma_x(x_n, \nabla m_n(x_n)), \frac{2\eta \tau (1 - \eta)}{B^2 C^2 \kappa_{\max}} \Gamma_x(x_n, \nabla m_n(x_n))^2 \right\}$$

$$= \Gamma_x(x_n, \nabla m_n(x_n)) \min \left\{ \frac{\eta \tau \Delta_n}{B}, \frac{2\eta \tau (1 - \eta)}{B^2 C^2 \kappa_{\max}} \Gamma_x(x_n, \nabla m_n(x_n))^2 \right\}$$

$$\geq \Gamma_x(x_n, \nabla m_n(x_n)) \min \left\{ \frac{\eta \tau}{B}, \frac{2\eta \tau (1 - \eta)}{B^2 C^2} \right\} \min \left\{ \Delta_n, 1, \frac{\Gamma_x(x_n, \nabla m_n(x_n))}{\kappa_{\max}} \right\}$$

That is, $x_n^r := h_n(\alpha^r_n)$ satisfies (201) for any

$$\mu_1 \in \left(0, \min \left\{ \frac{\eta \tau}{B}, \frac{2\eta \tau (1 - \eta)}{B^2 C^2} \right\} \right]$$

133
The proof of Lemma 7.22 shows that an Armijo step-size $\alpha_n^\Gamma$ can be obtained by setting $\alpha_n^\Gamma = \tau_{kn}$ where

$$k_n := \min \left\{ k \in \mathbb{N} : m_n(x_n) - m_n(h_n(\tau^k)) \geq -\eta \nabla m_n(x_n)^T (h_n(\tau^k) - x_n) \right\}$$

**Example 8.2.** Assume that $\mathcal{X}$ is nonempty, closed, and convex. Choose constant $\gamma > 0$ and $H \in S_{++}^{l \times l}$, and define the optimality function $\Gamma_{\mathcal{X}} : \mathcal{X} \times \mathbb{R}^l \mapsto \mathbb{R}$ by

$$\Gamma_{\mathcal{X}}(x, d) := \|\Pi_{\mathcal{X}}(x - \gamma H^{-1}d, H) - x\|_2$$

(213)

Next we show that $\Gamma_{\mathcal{X}}$ as defined in (213) is a valid optimality function.

**Lemma 8.6.** Suppose that the set $\mathcal{X}$ is nonempty, closed and convex. Then $\Gamma_{\mathcal{X}}$ as defined in (213) satisfies Properties P 8.1 and P 8.2.

**Proof.** Clearly $\Gamma_{\mathcal{X}}(x, d) \geq 0$ for all $x \in \mathcal{X}$ and $d \in \mathbb{R}^l$. It follows from Property P 7.6 that for any continuously differentiable function $g : \mathcal{X} \mapsto \mathbb{R}$ and any $x \in \mathcal{X}$, $\Gamma_{\mathcal{X}}(x, \nabla g(x)) = 0$ if and only if $\nabla g(x)^T (y - x) \geq 0$ for all $y \in \mathcal{X}$. Therefore, $\Gamma_{\mathcal{X}}$ satisfies Property P 8.1.

Consider any sequence $\{(x_n, d_n)\}_{n \in \mathbb{N}} \subset \mathcal{X} \times \mathbb{R}^l$ such that $(x_n, d_n) \to (x, d)$ for some $(x, d) \in \mathcal{X} \times \mathbb{R}^l$. Then it follows from the continuity of $\Pi_{\mathcal{X}}$ that

$$\lim_{n \to \infty} \Gamma_{\mathcal{X}}(x_n, d_n) = \lim_{n \to \infty} \|\Pi_{\mathcal{X}}(x_n - \gamma H^{-1}d_n, H) - x\|_2 = \|\Pi_{\mathcal{X}}(x - \gamma H^{-1}d, H) - x\|_2 = \Gamma_{\mathcal{X}}(x, d)$$

Thus, $\Gamma_{\mathcal{X}}$ is a continuous function and thus it satisfies Property P 8.2. \qed

Next we discuss how, given $x_n \in \mathcal{X} \cap T_n$ that is not a stationary point of $m_n$ in $\mathcal{X}$, one can find a point $x_n^\Gamma \in \mathcal{X} \cap T_n$ that satisfies (201) with respect to $\Gamma_{\mathcal{X}}$ defined above. As in the previous example, we define an arc $h_n : [0, 1] \mapsto \mathcal{X} \cap T_n$, find an Armijo step-size $\alpha_n^\Gamma$ that satisfies the Armijo sufficient conditions (206)–(208), and set $x_n^\Gamma := h_n(\alpha_n^\Gamma)$. First, choose a target point $\Phi_n \in \mathcal{X}$ as follows:

$$\Phi_n := \Pi_{\mathcal{X}}(x_n - \gamma H^{-1}\nabla m_n(x_n), H)$$

(214)

Then it follows from the definition of $\Gamma_{\mathcal{X}}$ in (213) that

$$\Gamma_{\mathcal{X}}(x_n, \nabla m_n(x_n)) = \|\Phi_n - x_n\|_2$$

(215)
Note that $\Phi_n = x_n$ if and only if $x_n$ is a stationary point of $m_n$ in $X$.

The arc $h_n$ is the same as in (205). If $x_n \in X \cap T_n$ is not a stationary point of $m_n$ in $X$, then

$$\|h_n(\alpha) - x_n\| = \alpha \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \|\Phi_n - x_n\|$$

$$= \alpha \min \{ \Delta_n, \|\Phi_n - x_n\| \} \leq \Delta_n$$

for any $\alpha \in [0, 1]$. Thus, $h_n(\alpha) \in T_n$ for all $\alpha \in [0, 1]$. Further, since $x_n, \Phi_n \in X$ and $X$ is convex, $h_n(\alpha) \in X$ for all $\alpha \in [0, 1]$. Hence $h_n(\alpha) \in X \cap T_n$ for all $\alpha \in [0, 1]$.

The following lemma shows that an Armijo step-size $\alpha^*_n$ that satisfies the Armijo sufficient conditions (206)–(208) exists for the arc $h_n$.

**Lemma 8.7.** Suppose that the set $X$ is nonempty, closed and convex, and that the model function $m_n : X \mapsto \mathbb{R}$ satisfies Assumption A 8.3. Consider any $x_n$ that is not a stationary point of $m_n$ in $X$. Let the target point $\Phi_n \in X$ be given by (214). Let the arc $h_n : [0, 1] \mapsto X \cap T_n$ be given by (205). Then there exists a step-size $\alpha^*_n \in [0, 1]$ that satisfies (206)–(208).

**Proof.** We establish the result by showing that all the conditions of Lemma 7.22 are satisfied. First, it follows from Assumption A 8.3 that $m_n$ is continuously differentiable. Since $x_n$ is not a stationary point of $m_n$ in $X$, $\Phi_n \neq x_n$, and hence

$$h_n(\alpha) = x_n + \alpha \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} (\Phi_n - x_n) \quad \text{for all } \alpha \in [0, 1]$$

Next we show that $h_n$ satisfies (166)–(168).

$$\lim_{\alpha \downarrow 0} \|h_n(\alpha) - x_n\|_2 = \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \|\Phi_n - x_n\|_2 \lim_{\alpha \downarrow 0} \alpha = 0$$

Thus, $h_n$ satisfies (166).

Since $x_n$ is not a stationary point of $m_n$ in $X$, it follows from Property P 7.7 that $\nabla m_n(x_n)^T (\Phi_n - x_n) < 0$. Hence

$$\nabla m_n(x_n)^T (h_n(\alpha) - x_n) = \alpha \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \nabla m_n(x_n)^T (\Phi_n - x_n) < 0$$

for all $\alpha \in (0, 1]$. Thus, $h_n$ satisfies (167).
Also,
\[
\limsup_{\alpha \to 0} \nabla m_n(x_n)^T \frac{h_n(\alpha) - x_n}{\|h_n(\alpha) - x_n\|_2} = \nabla m_n(x_n)^T \frac{\Phi_n - x_n}{\|\Phi_n - x_n\|_2} < 0
\]
Therefore, \(h_n\) satisfies (168). Hence, it follows from Lemma 7.22 that there exists \(\alpha^\Gamma_n \in [0, 1]\) that satisfies (206)—(208).

Next we show that \(x^\Gamma_n := h_n(\alpha^\Gamma_n)\), where \(\alpha^\Gamma_n\) satisfies the Armijo sufficient conditions, satisfies (201) for the optimality function \(\Gamma_x\).

**Lemma 8.8.** Suppose that the set \(\mathcal{X}\) is nonempty, closed and convex, and that the model function \(m_n : \mathcal{X} \mapsto \mathbb{R}\) satisfies Assumptions A 8.3 and A 8.4. Consider any \(x_n\) that is not a stationary point of \(m_n\) in \(\mathcal{X}\). Let the target point \(\Phi_n \in \mathcal{X}\) be given by (214). Let the arc \(h_n : [0, 1] \mapsto \mathcal{X} \cap T_n\) be given by (205), and let \(\alpha^\Gamma_n\) be any Armijo step-size that satisfies (206)–(208). Then \(x^\Gamma_n := h_n(\alpha^\Gamma_n)\) satisfies (201) with
\[
0 < \mu_1 < \min \left\{ \frac{\eta \lambda_{\min}(H)}{\gamma C}, \frac{2\eta(1-\eta)\tau \lambda_{\min}(H)^2}{\gamma^2} \right\}
\]
where \(\eta, \tau \in (0, 1)\) are the parameters associated with the Armijo step-size \(\alpha^\Gamma_n\), and \(C\) is the constant in (194).

**Proof.** Since \(\Phi_n := \Pi_\mathcal{X}(x_n - \gamma H^{-1} \nabla m_n(x_n), H)\), it follows from Property P 7.7 that
\[
-\nabla m_n(x_n)^T (\Phi_n - x_n) \geq \frac{\lambda_{\min}(H)}{\gamma} \|\Phi_n - x_n\|_2^2 \quad (216)
\]
Since \(x_n\) is not a stationary point of \(m_n\) in \(\mathcal{X}\), \(\Phi_n \neq x_n\), and hence
\[
h_n(\alpha) = x_n + \alpha \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} (\Phi_n - x_n) \quad \text{for all } \alpha \in [0, 1]
\]
Consider the following cases regarding \(\alpha^\Gamma_n\).

Case(a) Suppose that \(\alpha^\Gamma_n \in [\tau \bar{\alpha}_n, 1]\) and \(\bar{\alpha}_n \geq 1\). Then \(\alpha^\Gamma_n \geq \tau\), and hence it follows from (216),
\( (194) \) and \( (215) \) that
\[
\begin{align*}
m_n(x_n) - m_n(h_n(\alpha_n^r)) & \geq -\eta \nabla m_n(x_n)^T(h_n(\alpha_n^r) - x_n) \\
& = -\eta \alpha_n^r \min \left\{ \frac{\Delta_n}{\gamma} \left\| \Phi_n - x_n \right\|, 1 \right\} \nabla m_n(x_n)^T(\Phi_n - x_n) \\
& \geq \eta \alpha_n^r \min \left\{ \frac{\Delta_n}{\gamma} \left\| \Phi_n - x_n \right\|, 1 \right\} \lambda_{\text{min}}(H) \left\| \Phi_n - x_n \right\|_2^2 \\
& \geq \eta \tau \lambda_{\text{min}}(H) \min \left\{ \frac{\Delta_n}{\gamma} \left\| \Phi_n - x_n \right\|, 1 \right\} \left\| \Phi_n - x_n \right\|_2^2 \\
& \geq \eta \tau \lambda_{\text{min}}(H) \min \left\{ \frac{\Delta_n}{\gamma} \frac{\Gamma(x_n, \nabla m_n(x_n))}{\gamma_{\text{max}}} \right\} \Gamma(x_n, \nabla m_n(x_n)) \quad (217)
\end{align*}
\]
where the last inequality follows from \( C \geq 1 \) and \( \gamma_{\text{max}} \geq 1 \).

Case (b) Suppose that \( \alpha_n^r \in [\tau \alpha_n, 1] \), and \( \bar{\alpha}_n < 1 \) satisfies \( (208) \). For brevity of notation, let
\[
\hat{\alpha}_n := \bar{\alpha}_n \min \left\{ \frac{\Delta_n}{\gamma} \left\| \Phi_n - x_n \right\|, 1 \right\}
\]
Since \( m_n \) is twice continuously differentiable on \( \mathcal{X} \), there is \( y = x_n + t \hat{\alpha}_n(\Phi_n - x_n) \) for some \( t \in (0, 1) \) such that
\[
m_n(x_n + \hat{\alpha}_n(\Phi_n - x_n)) = m_n(x_n) + \hat{\alpha}_n \nabla m_n(x_n)^T(\Phi_n - x_n) + \frac{1}{2} \hat{\alpha}_n^2(\Phi_n - x_n)^T \nabla^2 m_n(y)(\Phi_n - x_n)
\]
Note that \( x_n + \hat{\alpha}_n(\Phi_n - x_n) = h_n(\bar{\alpha}_n) \in \mathcal{X} \cap \mathcal{T}_n \). Since \( x_n \in \mathcal{X} \cap \mathcal{T}_n \) and \( \mathcal{X} \cap \mathcal{T}_n \) is convex, it follows that \( y \in \mathcal{X} \cap \mathcal{T}_n \). It follows from Assumption A 8.4 that
\[
m_n(h_n(\bar{\alpha}_n)) - m_n(x_n) - \hat{\alpha}_n \nabla m_n(x_n)^T(\Phi_n - x_n) \leq \frac{1}{2} \hat{\alpha}_n^2 \gamma_{\text{max}} \left\| \Phi_n - x_n \right\|_2^2 \quad (218)
\]
Since \( \bar{\alpha}_n \) satisfies \( (208) \), it also holds that
\[
m_n(h_n(\bar{\alpha}_n)) - m_n(x_n) > \eta \bar{\alpha}_n \min \left\{ \frac{\Delta_n}{\gamma} \left\| \Phi_n - x_n \right\|, 1 \right\} \nabla m_n(x_n)^T(\Phi_n - x_n) = \eta \bar{\alpha}_n \nabla m_n(x_n)^T(\Phi_n - x_n) \quad (219)
\]
It follows from \( (218) \), \( (219) \), and \( (216) \) that
\[
\frac{1}{2} \hat{\alpha}_n^2 \gamma_{\text{max}} \left\| \Phi_n - x_n \right\|_2^2 > -(1 - \eta) \hat{\alpha}_n \nabla m_n(x_n)^T(\Phi_n - x_n) \geq \frac{(1 - \eta) \lambda_{\text{min}}(H)}{\gamma} \hat{\alpha}_n \left\| \Phi_n - x_n \right\|_2^2
\]
\[
\Rightarrow \quad \hat{\alpha}_n > \frac{2(1 - \eta) \lambda_{\text{min}}(H)}{\gamma \gamma_{\text{max}}}
\]
137
It follows from the Armijo condition (206), (216) and (215) that

\[
m_n(x_n) - m_n(h_n(\alpha_n^r)) \geq -\eta \nabla m_n(x_n)^T (h_n(\alpha_n^r) - x_n) \quad (220) \\
= -\eta \alpha_n^r \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \nabla m_n(x_n)^T (\Phi_n - x_n) \quad (221) \\
\geq \eta \alpha_n^r \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \frac{\lambda_{\min}(H)}{\gamma} \|\Phi_n - x_n\|^2 \quad (222) \\
\geq \eta \tau \alpha_n \min \left\{ \frac{\Delta_n}{\|\Phi_n - x_n\|}, 1 \right\} \frac{\lambda_{\min}(H)}{\gamma} \|\Phi_n - x_n\|^2 \quad (223) \\
= \frac{\eta \tau \lambda_{\min}(H)}{\gamma} \hat{\alpha}_n \|\Phi_n - x_n\|^2 \quad (224) \\
> \frac{\eta \tau \lambda_{\min}(H)}{\gamma} \frac{2(1 - \eta) \lambda_{\min}(H)}{\gamma \kappa_{\max}} \|\Phi_n - x_n\|^2 \quad (225) \\
= \frac{2\eta(1 - \eta) \tau \lambda_{\min}(H)^2}{\gamma^2 \kappa_{\max}} \Gamma_x(x_n, \nabla m_n(x_n))^2 \quad (226)
\]

Combining (217) and (226) it follows that

\[
m_n(x_n) - m_n(h_n(\alpha_n^r)) \geq \min \left\{ \frac{\eta \tau \lambda_{\min}(H)}{\gamma C} \min \left\{ \frac{\Delta_n}{\kappa_{\max}}, \frac{2\eta(1 - \eta) \tau \lambda_{\min}(H)^2}{\gamma^2 \kappa_{\max}} \right\} \Gamma_x(x_n, \nabla m_n(x_n)), \right. \\
\left. \frac{\eta \tau \lambda_{\min}(H)}{\gamma C}, \frac{2\eta(1 - \eta) \tau \lambda_{\min}(H)^2}{\gamma^2 \kappa_{\max}} \right\} \Gamma_x(x_n, \nabla m_n(x_n))^2 \min \left\{ 1, \frac{\Gamma_x(x_n, \nabla m_n(x_n))}{\kappa_{\max}} \right\} 
\]

That is, \(x_n^r := h_n(\alpha_n^r)\) satisfies (201) for any

\[
\mu_1 \in \left( 0, \min \left\{ \frac{\eta \tau \lambda_{\min}(H)}{\gamma C}, \frac{2\eta(1 - \eta) \tau \lambda_{\min}(H)^2}{\gamma^2 \kappa_{\max}} \right\} \right]
\]

As for the previous example, the proof of Lemma 7.22 shows that an Armijo step-size \(\alpha_n^r\) can be obtained by setting \(\alpha_n^r = \tau^{k_n}\) where

\[
k_n := \min \left\{ k \in \mathbb{N} : m_n(x_n) - m_n(h_n(\tau^k)) \geq -\eta \nabla m_n(x_n)^T (h_n(\tau^k) - x_n) \right\}
\]

We refer the reader to ? for further examples of widely used optimality functions and model function for which the existence of a point \(x^r_n\) satisfying Assumption 8.5 can be demonstrated.

Finally, we would like to note that for any pair \(\Gamma_x\) and \(m_n\) for which a point \(x^r_n\) exists satisfying (201), we can choose not only \(x_n + d_n = x^r_n\) in order to satisfy (200) but also \(x_n + d_n = x^*_n\) for
any \( x^*_n \in \arg \min_{x \in \mathcal{X} \cap \mathcal{T}_n} m_n(x) \). This is a valid choice since for any \( x^*_n \in \arg \min_{x \in \mathcal{X} \cap \mathcal{T}_n} m_n(x) \), we have

\[
m_n(x_n) - m_n(x^*_n) \geq m_n(x_n) - m_n(x^*_n)
\]  

(227)

Next, we establish the convergence of Algorithm 8.2. We will use the following notation. In Algorithm 8.2, if \( x_n \) is a stationary point of the model function \( m_n \) in \( \mathcal{X} \), then iteration \( n \) is called a null iteration. Otherwise, if \( \rho_n \geq \eta_1 \), then iteration \( n \) is called a successful iteration and if not, it is called an unsuccessful iteration. Let

\[
\mathcal{N} := \{ n \in \mathbb{N} : \nabla m_n(x_n)^T (x - x_n) \geq 0 \quad \forall \ x \in \mathcal{X} \} \\
\mathcal{K} := \{ n \in \mathbb{N} : \rho_n \geq \eta_1 \} \\
\mathcal{U} := \{ n \in \mathbb{N} : \rho_n < \eta_1 \}
\]

Lemma 8.9. Consider a sequence \( \{ \Delta_n \}_{n \in \mathbb{N}} \) of trust-region radii generated by Algorithm 8.2. If

\[
\lim_{n \to \infty, n \in \mathcal{K}} \Delta_n = 0
\]

then

\[
\lim_{n \to \infty} \Delta_n = 0
\]

Proof. First, consider the case in which there is \( N_1 \in \mathbb{N} \) such that \( n \in \mathcal{K} \) for all \( n > N_1 \). Then it immediately follows from the assumption that \( \lim_{n \to \infty, n \in \mathcal{K}} \Delta_n = 0 \), that \( \lim_{n \to \infty} \Delta_n = 0 \).

Otherwise, consider the subsequences \( \mathcal{K}^+ := \{ n + 1 \in \mathbb{N} : n \in \mathcal{K} \} \) and \( \mathcal{N} \cup \mathcal{U} \setminus \mathcal{K}^+ \). Note that \( \mathcal{K} \cup \mathcal{K}^+ \cup (\mathcal{N} \cup \mathcal{U} \setminus \mathcal{K}^+) = \mathbb{N} \) and that \( |\mathcal{K}^+| = |\mathcal{K}| = \infty \). For each \( n \in \mathcal{K} \), \( \Delta_{n+1} \leq \sigma_3 \Delta_n \). Thus

\[
\lim_{n \to \infty, n \in \mathcal{K}^+} \Delta_n = \lim_{n \to \infty, n \in \mathcal{K}} \Delta_{n+1} \leq \lim_{n \to \infty, n \in \mathcal{K}} \sigma_3 \Delta_n = 0
\]

and hence \( \lim_{n \to \infty, n \in \mathcal{K}^+} \Delta_n = 0 \).

For each \( n \in \mathcal{N} \cup \mathcal{U} \setminus \mathcal{K}^+ \), let \( N(n) := \max\{ n' \in \mathcal{K}^+ : n' < n \} \). For each \( n \in \mathcal{N} \cup \mathcal{U} \setminus \mathcal{K}^+ \), \( \Delta_n \leq \Delta_{N(n)} \). Note that since \( |\mathcal{K}^+| = \infty \), \( N(n) \to \infty \) as \( n \to \infty \). Thus

\[
\lim_{n \to \infty, n \in \mathcal{N} \cup \mathcal{U} \setminus \mathcal{K}^+} \Delta_n \leq \lim_{n \to \infty, n \in \mathcal{N} \cup \mathcal{U} \setminus \mathcal{K}^+} \Delta_{N(n)} = \lim_{n \to \infty, n \in \mathcal{K}^+} \Delta_n = 0
\]

and hence \( \lim_{n \to \infty, n \in \mathcal{N} \cup \mathcal{U} \setminus \mathcal{K}^+} \Delta_n = 0 \). Therefore

\[
\lim_{n \to \infty} \Delta_n = 0
\]
Lemma 8.10. Suppose that the set $\mathcal{X}$ is nonempty and convex, and that $f$ is continuously differentiable on a neighborhood of $\mathcal{X}$. Suppose that the sequence $\{f_n\}_{n \in \mathbb{N}}$ satisfies Assumption A 8.1 and that the sequence $\{\nabla_n f\}_{n \in \mathbb{N}}$ satisfies Assumption A 8.2. Suppose that the sequence $\{m_n\}_{n \in \mathbb{N}}$ of model functions satisfies Assumptions A 8.3 and A 8.4. Suppose that the sequences $\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X}$, $\{x_n + d_n\}_{n \in \mathbb{N}} \subset \mathcal{X}$, and $\{\Delta_n\}_{n \in \mathbb{N}} \subset \mathbb{R}$ satisfy $\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D}$ where $\mathcal{D}$ is a compact subset of $\mathcal{X}$, $\|d_n\| \leq \Delta_n$ for all $n$, and $\Delta_n \to 0$ as $n \to \infty$. Then

$$
\lim_{n \to \infty} \frac{|f_n(x_n + d_n) - m_n(x_n + d_n)|}{\Delta_n} = 0
$$

(228)

Proof. Since $\mathcal{X}$ is convex, $[x_n, x_n + d_n] \subset \mathcal{X}$. Since $m_n$ satisfies Assumption A 8.3, there is $s_n \in (0, 1)$ such that $H_n := \nabla^2 m_n(x_n + s_n d_n)$ satisfies

$$
m_n(x_n + d_n) = f_n(x_n) + \nabla_n f(x_n)^T d_n + \frac{1}{2} d_n^T H_n d_n
$$

Also, since $f$ is continuously differentiable on $\mathcal{X}$, there is $t_n \in (0, 1)$ such that

$$
f(x_n + d_n) - f(x_n) = \nabla f(x_n + t_n d_n)^T d_n
$$

Let

$$
a_n := \begin{cases} 
1 & \text{if } d_n = 0 \\
0 & \text{otherwise}
\end{cases}
$$

Thus,

$$
|f_n(x_n + d_n) - m_n(x_n + d_n)| = \left| f_n(x_n + d_n) - f_n(x_n) - \nabla_n f(x_n)^T d_n - \frac{1}{2} d_n^T H_n d_n \right|
$$

$$
\leq |(f_n - f)(x_n + d_n) - (f_n - f)(x_n)| + |f(x_n + d_n) - f(x_n) - \nabla f(x_n)^T d_n| + \frac{1}{2} d_n^T H_n d_n
$$

$$
\leq \frac{|(f_n - f)(x_n + d_n) - (f_n - f)(x_n)|}{\|d_n\|_2 + a_n} \|d_n\|_2 + \frac{1}{2} d_n^T H_n d_n
$$

$$
\leq \frac{|(f_n - f)(x_n + d_n) - (f_n - f)(x_n)|}{\|d_n\|_2 + a_n} \|d_n\|_2 + \frac{1}{2} d_n^T H_n d_n
$$

$$
\leq \frac{|(f_n - f)(x_n + d_n) - (f_n - f)(x_n)|}{\|d_n\|_2 + a_n} \|d_n\|_2 + \frac{1}{2} d_n^T H_n d_n
$$

$$
+ \|\nabla f(x_n + t_n d_n) - \nabla f(x_n)\|_2 \|d_n\|_2
$$

$$
+ \|\nabla f(x_n) - \nabla_n f(x_n)\|_2 \|d_n\|_2 + \frac{1}{2} \|H_n\|_2 \|d_n\|^2
$$

140
Let

\[ A_n := \frac{|(f_n - f)(x_n + d_n) - (f_n - f)(x_n)|}{\|d_n\|^2 + a_n} \quad (229) \]

\[ B_n := \|\nabla f(x_n + t_n d_n) - \nabla f(x_n)\|_2 \quad (230) \]

\[ C_n := \left\| \nabla f(x_n) - \nabla f(x_n) \right\|_2 \quad (231) \]

It follows from (194) and \( \|d_n\| \leq \Delta_n \) that \( \|d_n\|_2 \leq C \|d_n\| \leq C \Delta_n \). Also, since \( m_n \) satisfies Assumption A 8.4, \( \|H_n\|_2 \leq \kappa_n \leq \kappa_{\text{max}} \). Thus,

\[ |f_n(x_n + d_n) - m_n(x_n + d_n)| \leq (A_n + B_n + C_n) \Delta_n + \frac{1}{2} \kappa_{\text{max}} C^2 \Delta_n^2 \]

and hence

\[ \frac{|f_n(x_n + d_n) - m_n(x_n + d_n)|}{\Delta_n} \leq (A_n + B_n + C_n) + \frac{1}{2} \kappa_{\text{max}} C^2 \Delta_n \quad (232) \]

Recall that \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D} \subset \mathcal{X} \) where \( \mathcal{D} \) is compact, \( \|d_n\| \leq \Delta_n \) for all \( n \), and \( \Delta_n \to 0 \) as \( n \to \infty \). It follows that \( \{x_n + d_n\}_{n \in \mathbb{N}} \) is bounded. Let \( \mathcal{A} \) denote the set of accumulation points of \( \{x_n + d_n\}_{n \in \mathbb{N}} \). Next we show by contradiction that \( \mathcal{A} \subset \mathcal{D} \). Suppose that there is \( x \in \mathcal{A} \setminus \mathcal{D} \).

Let \( \delta_{\mathcal{D}} := \min\{\|y - x\|_2 : y \in \mathcal{D}\} > 0 \). Since \( \Delta_n \to 0 \), there is \( N \in \mathbb{N} \) such that \( \Delta_n < \delta_{\mathcal{D}}/2 \) for all \( n > N \). Since \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D} \), \( \|x_n + d_n - x\|_2 \geq \|x_n - x\|_2 - \|d_n\|_2 \geq \min\{\|y - x\|_2 : y \in \mathcal{D}\} - \Delta_n > \delta_{\mathcal{D}}/2 \) for all \( n > N \), which contradicts \( x \in \mathcal{A} \). Thus \( \mathcal{A} \subset \mathcal{D} \subset \mathcal{X} \). Then it follows from Lemma 2.2 that \( \overline{\text{cl}}(\bigcup_{n \in \mathbb{N}} \{x_n + d_n\}) \) is a compact subset of \( \mathcal{X} \). Since \( \mathcal{X} \) is convex, it follows that \( \mathcal{D}' := \text{conv}(\mathcal{D} \cup \overline{\text{cl}}(\bigcup_{n \in \mathbb{N}} \{x_n + d_n\})) \) is a compact convex subset of \( \mathcal{X} \).

Since \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{D}' \subset \mathcal{X} \) and \( \mathcal{D}' \) is compact, it follows from Assumption A 8.1 that \( \|f_n - f\|_{\mathcal{W}_0(\mathcal{D}')} \to 0 \) as \( n \to \infty \). In particular,

\[ \lim_{n \to \infty} \sup_{\{x,y \in \mathcal{D}' : x \neq y\}} \frac{|(f_n - f)(y) - (f_n - f)(x)|}{\|y - x\|_2} = 0 \]

For each \( n \in \mathbb{N} \) such that \( d_n \neq 0 \),

\[ A_n = \left| \frac{(f_n - f)(x_n + d_n) - (f_n - f)(x_n)}{\|d_n\|^2} \right| \leq \sup_{\{x,y \in \mathcal{D}' : x \neq y\}} \frac{|(f_n - f)(y) - (f_n - f)(x)|}{\|y - x\|_2} \]

and for each \( n \in \mathbb{N} \) such that \( d_n = 0 \), \( A_n = 0 \). Therefore,

\[ \lim_{n \to \infty} A_n = 0 \]

Since \( x_n \in \mathcal{D}' \), \( x_n + d_n \in \mathcal{D}' \), and \( \mathcal{D}' \) is convex, it follows that \( x_n + t_n d_n \in \mathcal{D}' \). Since \( \nabla f \) is continuous on \( \mathcal{X} \) and \( \mathcal{D}' \subset \mathcal{X} \) is compact, \( \nabla f \) is uniformly continuous on \( \mathcal{D}' \). That is, for any \( \varepsilon > 0 \),
there is \( \delta > 0 \) such that

\[
\sup_{\{x, y \in D' : \|y - x\|_2 < \delta\}} \|\nabla f(y) - \nabla f(x)\|_2 < \varepsilon
\]

Since \( t_n \in [0, 1] \) and \( \Delta_n \to 0 \) as \( n \to \infty \), it follows that \( \|(x_n + t_n d_n) - x_n\| = t_n \|d_n\| \leq t_n \Delta_n \to 0 \) as \( n \to \infty \). Thus, there is \( N \in \mathbb{N} \) such that for all \( n > N \), \( \|x_n + t_n d_n - x_n\| < \delta \) and consequently

\[
\|\nabla f(x_n + t_n d_n) - \nabla f(x_n)\|_2 \leq \sup_{\{x, y \in D' : \|y - x\|_2 < \delta\}} \|\nabla f(y) - \nabla f(x)\|_2 < \varepsilon
\]

Therefore,

\[
\lim_{n \to \infty} B_n = \lim_{n \to \infty} \|\nabla f(x_n + t_n d_n) - \nabla f(x_n)\|_2 = 0
\]

Since \( f \) is continuously differentiable on \( X \), \( \{\hat{\nabla}_n f\}_{n \in \mathbb{N}} \) satisfies Assumption A 8.2, and \( \{x_n\}_{n \in \mathbb{N}} \) is contained in a compact subset of \( X \), it follows from Lemma 5.1 that

\[
\lim_{n \to \infty} C_n = \lim_{n \to \infty} \|\nabla f(x_n) - \hat{\nabla}_n f(x_n)\|_2 = 0
\]

Also, \( \Delta_n \to 0 \) as \( n \to \infty \). Combining the results above, it follows that

\[
\lim_{n \to \infty} \left[ (A_n + B_n + C_n)C + \frac{1}{2} \kappa_{\max} C^2 \Delta_n \right] = 0
\]

Therefore, it follows from (232) that

\[
\lim_{n \to \infty} \frac{|f_n(x_n + d_n) - m_n(x_n + d_n)|}{\Delta_n} = 0
\]

\[\square\]

**Theorem 8.11.** Suppose that the set \( X \) is nonempty and convex, and that \( f \) is continuously differentiable on a neighborhood of \( X \) and satisfies Assumption A 7.10. Suppose that the sequence \( \{f_n\}_{n \in \mathbb{N}} \) satisfies Assumption A 8.1 and that the sequence \( \{\hat{\nabla}_n f\}_{n \in \mathbb{N}} \) satisfies Assumption A 8.2. Suppose that the sequence \( \{m_n\}_{n \in \mathbb{N}} \) of model functions satisfies Assumptions A 8.3 and A 8.4. Suppose that the optimality function \( \Gamma_X \) and the model functions \( m_n \) together satisfy Assumption A 8.5. Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset X \) generated by Algorithm 8.2. Suppose that \( \{x_n\}_{n \in \mathbb{N}} \subset D \subset X \) where \( D \) is compact. Then,

\[
\liminf_{n \to \infty} \Gamma_X(x_n, \hat{\nabla}_n f(x_n)) = 0
\]
Proof. Since $\nabla m_n(x_n) = \hat{\nabla}_n f(x_n)$ for each $n \in \mathbb{N}$, it is sufficient to show that

$$\lim_{n \to \infty} \Gamma_x(x_n, \nabla m_n(x_n)) = 0$$

We establish the result by contradiction. Suppose that there exists $\varepsilon > 0$ and $M \in \mathbb{N}$ such that

$$\Gamma_x(x_n, \nabla m_n(x_n)) \geq \varepsilon$$

for all $n > M$.

First, note that if $n \in \mathbb{N}$, that is, $\nabla m_n(x_n) T(x-x_n) \geq 0$ for all $x \in X$, then $\Gamma_x(x_n, \nabla m_n(x_n)) = 0$. Thus, if (233) holds, then $n \notin \mathcal{N}$.

Next we show that if (233) holds, then

$$\lim_{n \to \infty} \Delta_n = 0$$

Case(a) Suppose that $|\mathcal{K}| < \infty$. Then there is $N \in \mathbb{N}$ such that $N \geq M$ and for all $n > N$, $\rho_n < \eta_1$. Then $\Delta_{n+1} \leq \sigma_2 \Delta_n$ for all $n > N$, and thus $\Delta_n \to 0$ as $n \to \infty$.

Case(b) Suppose that $|\mathcal{K}| = \infty$. First we show that \{f(x_n)\}$_{n \in \mathbb{N}}$ is a non-increasing sequence for $n$ large enough. Recall that if $n \in \mathcal{U}$, then Algorithm 8.2 sets $x_{n+1} = x_n$ and hence $f(x_{n+1}) = f(x_n)$. Next, consider the case when $n \in \mathcal{K}$, that is, $\rho_n \geq \eta_1$. Also, the point $x_{n+1} = x_n + d_n$ satisfies (200), and $\|d_n\| > 0$. Combining these and (233), it follows that for all $n > M$ such that $n \in \mathcal{K}$,

$$f_n(x_n) - f_n(x_{n+1}) \geq \eta_1 \mu_1 \varepsilon \min \left\{1, \Delta_n, \frac{\varepsilon}{\kappa_{\max}} \right\}$$

(234)

Also, for each $n \in \mathcal{K}$, since $\|d_n\| \leq C \|d_n\| \leq C \Delta_n$,

$$\begin{align*}
f_n(x_n) - f_n(x_{n+1}) &= f(x_n) - f(x_{n+1}) + [(f_n - f)(x_n) - (f_n - f)(x_{n+1})] \\
&\leq f(x_n) - f(x_{n+1}) + \frac{|(f_n - f)(x_n) - (f_n - f)(x_{n+1})| \|d_n\|}{\|d_n\|} \\
&\leq f(x_n) - f(x_{n+1}) + A_n C \Delta_n
\end{align*}$$

(235)

where

$$A_n := \frac{|(f_n - f)(x_n) - (f_n - f)(x_{n+1})| \|d_n\|}{\|d_n\|} \leq \sup_{\{x,y \in \mathcal{D} : x \neq y\}} \frac{|(f_n - f)(y) - (f_n - f)(x)|}{\|y - x\|}$$

143
It follows from Assumption A 8.1 and (196) that
\[
\lim_{n \to \infty} \sup_{\{x, y \in D : x \neq y\}} \frac{|(f_n - f)(y) - (f_n - f)(x)|}{\|y - x\|_2} = 0
\]
and thus
\[
\lim_{n \to \infty, n \in \mathcal{K}} A_n = 0
\]
Thus, there is \( N_A \in \mathbb{N} \) such that for all \( n > N_A \) and \( n \in \mathcal{K} \),
\[
A_n < \frac{\eta_1 \mu_1 \varepsilon}{2C} \min \left\{ \frac{1}{\Delta_{\text{max}}}, 1, \frac{\varepsilon}{\Delta_{\text{max}} \kappa_{\text{max}}} \right\} \tag{236}
\]
Combining (234) and (235), it follows that for all \( n > \max\{M, N_A\} \) such that \( n \in \mathcal{K} \),
\[
f(x_n) - f(x_{n+1}) \geq \eta_1 \mu_1 \varepsilon \min \left\{ 1, \Delta_n, \frac{\varepsilon}{\kappa_{\text{max}}} \right\} - CA_n \Delta_n
\]
\[
> \eta_1 \mu_1 \varepsilon \min \left\{ 1, \Delta_n, \frac{\varepsilon}{\kappa_{\text{max}}} \right\} - \frac{\eta_1 \mu_1 \varepsilon}{2} \min \left\{ \frac{\Delta_n}{\Delta_{\text{max}}}, \Delta_n, \frac{\varepsilon \Delta_n}{\Delta_{\text{max}} \kappa_{\text{max}}} \right\}
\]
\[
\geq \frac{\eta_1 \mu_1 \varepsilon}{2} \min \left\{ 1, \Delta_n, \frac{\varepsilon}{\kappa_{\text{max}}} \right\} > 0 \tag{237}
\]
Therefore, \( f(x_{n+1}) \leq f(x_n) \) for all \( n > \max\{M, N_A\} \).

Since, from Assumption A 7.10, \( f \) is bounded below on \( \mathcal{X} \), \( f(x_n) \to f^* > -\infty \). Then
\[
\sum_{n=1}^{\infty} [f(x_n) - f(x_{n+1})] = \lim_{N \to \infty} [f(x_1) - f(x_N)] = f(x_1) - f^* < \infty \tag{238}
\]
and thus
\[
\lim_{n \to \infty} [f(x_n) - f(x_{n+1})] = 0
\]
Thus it follows from (237) that
\[
\lim_{n \to \infty, n \in \mathcal{K}} \min \left\{ 1, \Delta_n, \frac{\varepsilon}{\kappa_{\text{max}}} \right\} = 0
\]
and hence
\[
\lim_{n \to \infty, n \in \mathcal{K}} \Delta_n = 0
\]
Therefore it follows from Lemma 8.9 that
\[
\lim_{n \to \infty} \Delta_n = 0
\]
Recall that for all \( n > M \) it holds that \( n \notin \mathcal{N} \), and thus it follows from (200), and (233) that
\[
m_n(x_n) - m_n(x_n + d_n) \geq \mu_1 \varepsilon \min \left\{ 1, \Delta_n, \frac{\varepsilon}{\kappa_{\text{max}}} \right\}
\]
Since $\Delta_n \to 0$ as $n \to \infty$, there is $N_1 \geq M$ such that for all $n \geq N_1$, $\Delta_n = \min \left\{1, \Delta_n, \frac{\varepsilon}{\kappa_{\max}} \right\}$. Consequently, for $n > N_1$,

$$m_n(x_n) - m_n(x_n + d_n) \geq \mu_1 \varepsilon \Delta_n$$

Recall that $m_n(x_n) = f_n(x_n)$ for all $n \in \mathbb{N}$. Thus, for all $n > N_1$,

$$|1 - \rho_n| = \frac{|f_n(x_n + d_n) - m_n(x_n + d_n)|}{m_n(x_n) - m_n(x_n + d_n)} \leq \frac{1}{\mu_1 \varepsilon} \frac{|f_n(x_n + d_n) - m_n(x_n + d_n)|}{\Delta_n}$$

It also follows from Lemma 8.10 that

$$\lim_{n \to \infty} \frac{|f_n(x_n + d_n) - m_n(x_n + d_n)|}{\Delta_n} = 0$$

Thus there is $N_2 \geq N_1$ such that for all $n \geq N_2$, $\rho_n \geq \eta_1$. Therefore, for all $n \geq N_2$, it holds that $n \in \mathcal{K}$, and hence $\Delta_{n+1} \geq \Delta_n$. Thus, $\Delta_n \geq \Delta_{N_2} > 0$ for all $n \geq N_2$. This contradicts the previous conclusion that $\Delta_n \to 0$ as $n \to \infty$. Therefore

$$\lim \inf_{n \to \infty} \Gamma_X(x_n, \hat{\nabla} f(x_n)) = 0$$

Corollary 8.12. Suppose that the assumptions of Theorem 8.11 hold. Then there exists a subsequence $\{n_k\}_{k \in \mathbb{N}} \subset \mathbb{N}$ such that $x_{n_k} \to x^* \in \mathcal{S}$.

Proof. It follows from Theorem 8.11 that there is a subsequence $\{n_k\}_{k \in \mathbb{N}}$ such that

$$\lim_{k \to \infty} \Gamma_X(x_{n_k}, \hat{\nabla} f(x_{n_k})) = 0$$

Since $\{x_{n_k}\}_{k \in \mathbb{N}}$ remains in a compact set $\mathcal{D} \subset \mathcal{X}$, there is a further subsequence $\{m_k\}_{k \in \mathbb{N}} \subset \{n_k\}_{k \in \mathbb{N}}$ and $x^* \in \mathcal{D}$ such that

$$\lim_{k \to \infty} x_{m_k} = x^*$$

It follows from Assumption A 8.2 that

$$\lim_{k \to \infty} \hat{\nabla} f(x_{m_k}) = \nabla f(x^*)$$

Next it follows from Property P 8.2 that

$$\Gamma_X(x^*, \nabla f(x^*)) \leq \lim \inf_{k \to \infty} \Gamma_X(x_{m_k}, \nabla f(x_{m_k})) = \lim \inf_{k \to \infty} \Gamma_X(x_{m_k}, \hat{\nabla} f(x_{m_k})) = 0$$

Then it follows from Property P 8.1 that $x^* \in \mathcal{S}$. \qed
Lemma 8.13. Suppose that \( f \) is continuously differentiable on a neighborhood of \( X \). Suppose that the sequence \( \{f_n\}_{n\in\mathbb{N}} \) satisfies Assumption A 8.1. Suppose that the sequence \( \{m_n\}_{n\in\mathbb{N}} \) of model functions satisfies Assumptions A 8.3 and A 8.4. Suppose that the optimality function \( \Gamma_X \) and the model functions \( m_n \) together satisfy Assumption A 8.5. Consider a sequence \( \{x_n\}_{n\in\mathbb{N}} \subset X \) generated by Algorithm 8.2. Suppose that \( \{x_n\}_{n\in\mathbb{N}} \subset D \subset X \) where \( D \) is compact. Then, for any \( \varepsilon > 0 \), there exists \( N_1(\varepsilon) \in \mathbb{N} \) such that for all \( n > N_1(\varepsilon) \),

\[
f(x_{n+1}) < f(x_n) + \varepsilon
\]

Proof. If \( n \in \mathcal{N} \cup \mathcal{U} \), then \( x_{n+1} = x_n \), and thus \( f(x_{n+1}) = f(x_n) \).

Suppose that \( n \in \mathcal{K} \). Then \( \rho_n \geq \eta_1 \) and \( x_{n+1} = x_n + d_n \). Also, if follows from Assumption A 8.5, and (200) that

\[
m_n(x_n) - m_n(x_n + d_n) \geq \mu_1 \Gamma_X(x_n, \nabla m_n(x_n)) \min \left\{ 1, \Delta_n, \frac{\Gamma_X(x_n, \nabla m_n(x_n))}{\kappa_{\max}} \right\}
\]

Thus, from the definition of \( \rho_n \),

\[
f_n(x_n) - f_n(x_{n+1}) \geq \eta_1 \mu_1 \Gamma_X(x_n, \nabla m_n(x_n)) \min \left\{ 1, \Delta_n, \frac{\Gamma_X(x_n, \nabla m_n(x_n))}{\kappa_{\max}} \right\} \geq 0
\]

Since \( 0 < \|x_{n+1} - x_n\| = \|d_n\| \leq \Delta_n \),

\[
f(x_n) - f(x_{n+1}) \geq f_n(x_n) - f_n(x_{n+1}) - \frac{|(f_n - f)(x_{n+1}) - (f_n - f)(x_n)|}{\|x_{n+1} - x_n\|_2} C \Delta_n
\]

\[
\geq - \frac{|(f_n - f)(x_{n+1}) - (f_n - f)(x_n)|}{\|x_{n+1} - x_n\|_2} C \Delta_n
\]

Let

\[
A_n := \frac{|(f_n - f)(x_{n+1}) - (f_n - f)(x_n)|}{\|x_{n+1} - x_n\|_2}
\]

Recall that \( \{x_n\}_{n\in\mathbb{N}} \subset D \) where \( D \) is compact. Hence, it follows from Assumption A 8.1 and (196) that

\[
\lim_{n \to \infty} A_n \leq \lim_{n \to \infty} \sup_{x, y \in D : x \neq y} \frac{|(f_n - f)(y) - (f_n - f)(x)|}{\|y - x\|_2} = 0
\]

Thus, for any \( \varepsilon > 0 \), there is \( N_1(\varepsilon) \) such that for all \( n > N_1(\varepsilon) \), \( A_n < \varepsilon / (C \Delta_{\max}) \). Then, for all \( n > N_1(\varepsilon) \),

\[
f(x_n) - f(x_{n+1}) \geq -CA_n \Delta_n \geq -CA_n \Delta_{\max} > -\varepsilon
\]

\[\square\]
Lemma 8.14. Suppose that \( f \) is continuously differentiable on a neighborhood of \( X \). Suppose that the sequence \( \{ f_n \}_{n \in \mathbb{N}} \) satisfies Assumption A 8.1 and that the sequence \( \{ \hat{\nabla}_n f \}_{n \in \mathbb{N}} \) satisfies Assumption A 8.2. Suppose that the sequence \( \{ m_n \}_{n \in \mathbb{N}} \) of model functions satisfies Assumptions A 8.3 and A 8.4. Suppose that the optimality function \( \Gamma_X \) and the model functions \( m_n \) together satisfy Assumption A 8.5. Consider a sequence \( \{ x_n \}_{n \in \mathbb{N}} \subset X \) generated by Algorithm 8.2. Then, for any compact set \( D \subset X \setminus S \), there exists \( N_2(D) \in \mathbb{N} \), such that for all \( n > N_2(D) \), if \( x_n \in D \), then

\[ f(x_{n+1}) \leq f(x_n) \]

Proof. If there is \( N_1 \in \mathbb{N} \) such that for all \( n > N_1 \), \( x_n \in X \setminus D \), then the result trivially holds with \( N_2(D) = N_1 \).

Next, suppose that for any \( N' \in \mathbb{N} \), there is \( n > N' \) such that \( x_n \in D \). We first show by contradiction that there is \( \varepsilon > 0 \) and \( N \in \mathbb{N} \) such that for all \( n > N \) with \( x_n \in D \), it holds that

\[ \Gamma_X(x_n, \nabla m_n(x_n)) \geq \varepsilon \] (239)

Suppose that no such \( \varepsilon \) and \( N \) exist. Then there is a subsequence \( \{ x_{n_k} \}_{k \in \mathbb{N}} \subset D \) such that

\[ \lim_{k \to \infty} \Gamma_X(x_{n_k}, \nabla m_{n_k}(x_{n_k})) = 0 \]

Since \( D \) is compact, there is a further subsequence \( \{ m_k \}_{k \in \mathbb{N}} \subset \{ n_k \}_{k \in \mathbb{N}} \) such that \( x_{m_k} \to x^* \in D \). Recall from Assumption A 8.3 that \( \nabla m_n(x_n) = \hat{\nabla}_n f(x_n) \). Also, it follows from Assumption A 8.2 that \( \hat{\nabla}_{m_k} f(x_{m_k}) \to \nabla f(x^*) \). Thus it follows from Property P 8.2 that

\[ 0 = \liminf_{k \to \infty} \Gamma_X(x_{m_k}, \nabla m_{m_k}(x_{m_k})) = \liminf_{k \to \infty} \Gamma_X(x_{m_k}, \hat{\nabla}_{m_k} f(x_{m_k})) \geq \Gamma_X(x^*, \nabla f(x^*)) \]

Thus \( \Gamma_X(x^*, \nabla f(x^*)) = 0 \), and hence it follows from Property P 8.1 that \( x^* \in S \), which contradicts the assumption that \( D \subset X \setminus S \). Therefore, there is \( \varepsilon > 0 \) and \( N \in \mathbb{N} \) such that for all \( n > N \) with \( x_n \in D \), \( \Gamma_X(x_n, \nabla m_n(x_n)) \geq \varepsilon \).

Next, consider any \( n > N \) with \( x_n \in D \). If \( n \in \mathcal{N} \cup U \), then \( x_{n+1} = x_n \), and hence \( f(x_{n+1}) = f(x_n) \) and the required result holds. If \( n \in K \), then \( x_{n+1} = x_n + d_n \) such that (200) holds. It follows from \( \Gamma_X(x_n, \hat{\nabla}_n f(x_n)) \geq \varepsilon \) that

\[ m_n(x_n) - m_n(x_{n+1}) \geq \mu_1 \varepsilon \min \left\{ 1, \Delta_n, \frac{\varepsilon}{\kappa_{\max}} \right\} \]
Also, since \( n \in K \), it holds that \( \rho_n \geq \eta_1 \). Thus
\[
f_n(x_n) - f_n(x_{n+1}) \geq \eta_1 \mu_1 \varepsilon \min \left\{ 1, \frac{\varepsilon}{\kappa_{\max}} \right\}
\]
Also, \( 0 < \|x_{n+1} - x_n\|_2 \leq C \|d_n\| \leq C \Delta_n \). Hence
\[
f_n(x_n) - f_n(x_{n+1}) = f(x_n) - f(x_{n+1}) + [(f_n - f)(x_n) - (f_n - f)(x_{n+1})]
\leq f(x_n) - f(x_{n+1}) + A_n C \Delta_n
\]
where
\[
A_n := \frac{|(f_n - f)(x_{n+1}) - (f_n - f)(x_n)|}{\|d_n\|_2}
\]
It follows from Assumption A 8.1 that \( A_n \to 0 \) as \( n \to \infty \). Thus there is \( N_2(D) > N \) such that for all \( n > N_2(D) \),
\[
A_n < \frac{\eta_1 \mu_1 \varepsilon}{2C} \min \left\{ 1, \frac{1}{\Delta_{\max}}, \frac{\varepsilon}{\Delta_{\max} \kappa_{\max}} \right\}
\]
Thus, for all \( n > N_2(D) \) with \( n \in K \) and \( x_n \in D \),
\[
f(x_n) - f(x_{n+1}) \geq \eta_1 \mu_1 \varepsilon \min \left\{ 1, \frac{\varepsilon}{\kappa_{\max}} \right\} - C A_n \Delta_n
\geq \frac{\eta_1 \mu_1 \varepsilon}{2} \min \left\{ 1, \frac{\varepsilon}{\kappa_{\max}} \right\} > 0
\]
Therefore, \( f(x_{n+1}) \leq f(x_n) \) for all \( n > N_2(D) \) such that \( x_n \in D \).

Now we can use Theorem 6.1 to show that any sequence \( \{x_n\}_{n \in \mathbb{N}} \) generated by Algorithm 8.2 satisfies
\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in S} f(x)
\]

**Theorem 8.15.** Consider a function \( f : X \to \mathbb{R} \), a set \( S \subset X \), and a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset X \). Let \( A \) denote the set of accumulation points of \( \{x_n\}_{n \in \mathbb{N}} \) in \( X \). Suppose that \( \mu := \sup_{x \in S} f(x) < \infty \). (If \( \sup_{x \in S} f(x) = \infty \), then the result follows immediately.) Suppose that the following assumptions hold:

**A 8.6.** The function \( f \) is continuous on \( X \).

**A 8.7.** There exists an accumulation point of \( \{x_n\}_{n \in \mathbb{N}} \) that lies in \( \operatorname{cl}(S) \), i.e., \( A \cap \operatorname{cl}(S) \neq \emptyset \).

**A 8.8.** For any \( \varepsilon > 0 \), there exists \( N_1(\varepsilon) \in \mathbb{N} \) such that \( f(x_{n+1}) \leq f(x_n) + \varepsilon \) for all \( n \geq N_1(\varepsilon) \).

**A 8.9.** For any \( \varepsilon > 0 \), there exists \( N_2(\varepsilon) \in \mathbb{N} \) such that for all \( n \geq N_2(\varepsilon) \), if \( \mu + \varepsilon \leq f(x_n) \leq \mu + 2\varepsilon \), then \( f(x_{n+1}) \leq f(x_n) \).

148
Then
\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in S} f(x)
\] (240)

**Theorem 8.16.** Suppose that the set \( \mathcal{X} \) is nonempty, compact, and convex, and that \( f \) is continuously differentiable on a neighborhood of \( \mathcal{X} \). Suppose that the sequence \( \{f_n\}_{n \in \mathbb{N}} \) satisfies Assumption A 8.1 and that the sequence \( \{\nabla f_n\}_{n \in \mathbb{N}} \) satisfies Assumption A 8.2. Suppose that the sequence \( \{m_n\}_{n \in \mathbb{N}} \) of model functions satisfies Assumptions A 8.3 and A 8.4. Suppose that the optimality function \( \Gamma_X \) and the model functions \( m_n \) together satisfy Assumption A 8.5. Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) generated by Algorithm 8.2. Then,
\[
\limsup_{n \to \infty} f(x_n) \leq \sup_{x \in S} f(x)
\]

In addition, if Assumptions A 6.5 and A 6.6 are satisfied, then
\[
\lim_{n \to \infty} d(x_n, S) = 0
\]

**Proof.** We establish the result by showing that all the assumptions of Theorem 8.15 are satisfied. Corollary 8.12 establishes that Assumption A 6.2 holds. Lemma 8.13 establishes that Assumption A 6.3 holds. Lemma 8.14 establishes that Assumption A 6.4 holds. If Assumptions A 6.5 and A 6.6 are satisfied, then it follows from Corollary 6.4 that
\[
\lim_{n \to \infty} d(x_n, S) = 0
\]
\(\square\)