1 Introduction and Review of Related Literature

In this article we study the following stochastic optimization problem. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\zeta(\omega)$ (where $\omega$ denotes a generic element of $\Omega$) be a random variable on $(\Omega, \mathcal{F}, P)$, taking values in the probability space $(\Xi, \mathcal{G}, Q)$, where $Q$ denotes the probability measure of $\zeta$. Suppose for some open set $E \subset \mathbb{R}^l$, $F : E \times \Xi \rightarrow \mathbb{R}$ is a real-valued function such that for each $x \in E$, $F(x, \cdot) : \Xi \rightarrow \mathbb{R}$ is $\mathcal{G}$-measurable and integrable. Then, given a closed and convex set $X \subset E$, we consider the problem

$$\min_{x \in X} \left \{ f(x) := \mathbb{E}_Q [F(x, \zeta)] \right \} \quad \text{(P)}$$

If $f$ is sufficiently smooth on $X$ and its higher order derivatives can be computed exactly without much effort at any $x \in X$, then there are many deterministic optimization techniques that can be used to solve problem (P). However, in many practical problems, one is faced with the following conditions.

- For a given $x \in X$ and $\zeta \in \Xi$, $F(x, \zeta)$ can easily be computed exactly.
- It is prohibitively expensive to compute $f(x)$ or its higher order derivatives exactly at any $x \in X$, typically because it is difficult to compute the (often multidimensional) integral that defines $f(x)$.
- For a fixed $x \in X$ and $\zeta \in \Xi$, $\nabla_x F(x, \zeta)$ can not easily be computed exactly, or may not even exist at all $(x, \zeta)$, even though $f$ is differentiable at all $x$.

For example, one may have a simulator that takes $x$ as input, generates $\zeta$ according to a specified distribution, and computes $F(x, \zeta)$ with a single simulation run. Often such a simulator does not compute
∇_x F(x, ζ). This may be because the simulator code required to compute ∇_x F(x, ζ) is so complicated that the analyst does not want to take the time and effort to do the coding or run the risk of introducing errors in the code. It may also be that simulator is a black box to the user who wants to do the optimization and the simulator only returns F(x, ζ) for a given x, and the user cannot or does not want to change the simulator to also compute ∇_x F(x, ζ). Also, as mentioned above, it may be that ∇_x F(x, ζ) does not exist at all (x, ζ) even though f is differentiable at all x. A combination of these reasons held in an application that the authors worked on, and motivated the work in this paper.

Under the above assumptions it is natural to approximate f using a sample average function that can be obtained at a relatively low cost as follows. In order to show how this is done, consider for some N ∈ N, a sequence \{ζ^j(ω)\} \_j∈N of i.i.d. random functions on (Ω, F, P) taking values in Ξ such that ζ^j is F-measurable for each j ∈ N and P is such that Q is the measure induced on (Ξ, G) by ζ^1. Then, we define the sample average function ŵ_N : R^l × Ω → R as

\[ ŵ_N(x, ω) := \frac{1}{N} \sum_{j=1}^{N} F(x, ζ^j(ω)) \] (1)

Then for each x ∈ R^l, ŵ_N(x, ω) is an unbiased and consistent estimator of f(x). That is, E_P[ŵ_N(x, ω)] = f(x) and

\[ \lim_{N \to \infty} ŵ_N(x, ω) = f(x) \text{ for } P\text{-almost all } ω \]

The above result follows from the strong law of large numbers. Similarly, under some stronger conditions on the probability spaces and the differentiability of the function F(·,ζ) with respect to x, it is possible to show that the random function ∇ŵ_N : R^l × Ω → R^l defined as

\[ ∇ŵ_N(x, ω) := \frac{1}{N} \sum_{j=1}^{N} ∇_x F(x, ζ^j(ω)) \]

is an unbiased and consistent estimator of ∇f(x) for each x ∈ R^l. Given any x ∈ A and N ∈ N, in order to evaluate the sample average function for some fixed ¯ω ∈ Ω, we must be able to generate the independent replications \{ζ\_j = ζ(j)(ω) : j = 1, ..., N\}. In practice, ζ is usually a real-valued random vector and hence the set \{ζ\_j : j = 1, ..., N\} can be obtained by first generating a set \{ξ_1, ..., ξ_N\} of pseudo-random numbers that are independent and uniformly distributed on [0, 1] and subsequently applying an appropriate transformation to this set.

There are two essentially different ways in which such sampling techniques can be incorporated in algorithms that solve the problem (P). First, let us consider the class of algorithms popularly known as the stochastic approximation approach. A description of the simplest algorithm in this class (first proposed in ?) is sufficient to illustrate the basic sampling strategy. The algorithm essentially generates a sequence of \{x_n\} \_n∈N ⊂ A, starting at some x_0 ∈ A, by applying the following recursion.

\[ x_{n+1} = \Pi_A \left( x_n - α_n ∇ŵ(x_n) \right) \] (2)
In (2), $\Pi_X$ denotes the projection operation onto the set $\mathcal{X}$ and $\nabla \hat{f}(x_n)$ is an estimate of $\nabla f(x_n)$ which is obtained for each iteration $n$ as follows. For some sample size $N \in \mathbb{N}$, we first generate $N$ independent replications $\{\tilde{\zeta}_j : j = 1, \ldots, N\}$ of the random vector $\zeta$, that are also independent of all replications generated in earlier iterations. Then, we set

$$\nabla \hat{f}(x_n) := \frac{1}{N} \sum_{j=1}^{N} \nabla_x F(x_n, \tilde{\zeta}_n)$$

Further, the sequence $\{\alpha_n\}_{n \in \mathbb{N}}$ of positive step-sizes is chosen to satisfy

$$\sum_{n=1}^{\infty} \alpha_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty$$

Under certain regularity conditions, it can be shown that the sequence $\{x_n\}_{n \in \mathbb{N}}$ (which is a sequence of random vectors) converges with probability one, to a local minimum of $f$ in $\mathcal{X}$. Such an approach is attractive because of its relative ease of implementation. However, in practice this approach has been found to be extremely sensitive to the choice of the sequence $\{\alpha_n\}_{n \in \mathbb{N}}$ of step-sizes; a bad choice of which can lead to an extremely slow rate of convergence. Therefore, subsequent work in this area has aimed to improve the rate of convergence of this algorithm by averaging the gradient estimates obtained in consecutive iterations along with the use of adaptive step-size sequences. Further, the algorithm has also been applied to non-smooth problems where an estimate of a sub-gradient of the objective function is obtained through sampling at each iteration. We refer the reader to ?, (?) and (?) for details. Also, ? and ? contain detailed treatments of stochastic approximation algorithms for various cases.

The second approach, known variously as stochastic counterpart method, sample path optimization and sample average approximation, works as follows. We first fix some $\bar{\omega} \in \Omega$ and some $N \in \mathbb{N}$, and generate $N$ independent replications of the random vector $\zeta$ denoted by $\{\tilde{\zeta}_j = \zeta^j(\bar{\omega}) : j = 1, \ldots, N\}$. Then, using this fixed sample, we solve the optimization problem

$$\min_{x \in \mathcal{X}} \hat{f}_N(x, \bar{\omega}) := \frac{1}{N} \sum_{j=1}^{N} F(x, \tilde{\zeta}_j)$$

Now, it is easily seen that $\hat{f}_N(\cdot, \bar{\omega})$ is a deterministic function of $x$, we can compute $\hat{f}_N(x, \bar{\omega})$ and its higher order derivatives (if they exist) exactly for any $x \in \mathcal{X}$. Therefore, the problem ($\hat{P}$) can be solved using any appropriate deterministic optimization algorithm. This is certainly an appealing feature since there exist a vast store of algorithms developed for deterministic optimization from which this choice can be made. Suppose that

$$v^*_N(\omega) := \min_{x \in \mathcal{X}} \hat{f}_N(x, \omega) \quad \text{and} \quad x^*_N(\omega) \in \arg\min_{x \in \mathcal{X}} \hat{f}_N(x, \omega),$$

Then, it can be shown under mild conditions that as $N \to \infty$, $v^*_N$ and $x^*_N$ converge respectively to the optimal objective value and optimal solution of the “true” problem ($P$), for $\mathbb{P}$-almost all $\omega$. Further, there is a well-developed theory of statistical inference for the optimal value $v^*_N$ and optimal solution $x^*_N$ of the problem ($\hat{P}$)
that helps in setting the sample size $N$ and in the design of stopping tests. In particular, given any candidate optimal solution $x^* \in \mathcal{X}$ for the true problem, ? shows how the optimality gap $f(x^*) - \min_{x \in \mathcal{X}} f(x)$ can be estimated by solving $M$ sample average problems as in $(\hat{P})$ for $M$ independent realizations $\{\omega_1, \ldots, \omega_M\}$. The same article also provides methods to statistically test the validity of the first order Karush-Kuhn-Tucker optimality conditions for the point $x^*$ using an independently generated estimate of the gradient $\nabla f(x^*)$.

We refer the reader to ? and ? for details regarding the stochastic counterpart method and to ? and ? for a general description of Monte Carlo sampling methods in the solution of stochastic programs.

In this paper, we deal with a specific class of stochastic programs that satisfy, apart from the previously stated assumptions regarding the intractability of computing $f$ or its higher order derivatives, the following assumptions.

A 1. (a) The cost of computing $F(x, \zeta)$ for a given $x \in \mathbb{R}^l$ and $\zeta \in \Xi$, while being small relative to the cost of computing $f(x)$, is still large enough to warrant attempts to lower the number of evaluations of $F(x, \zeta)$ as much as possible.

(b) No sensitivity measures related to $F$ can be computed directly. That is, if $F(\cdot, \zeta)$ is continuously differentiable with respect to $x$ for some $\zeta \in \Xi$, then we assume that $\nabla_x F(x, \zeta)$ cannot be evaluated for any $x \in \mathcal{X}$. Otherwise, if $F(\cdot, \zeta)$ is convex for some $\zeta \in \Xi$, then we assume that a subgradient cannot be computed for any $x \in \mathcal{X}$.

(c) The function $f$ is continuously differentiable on $\mathcal{E}$.

Indeed, the first two statements of Assumption A 1 hold quite often in the case of simulation optimization. In many practical settings, $F(x, \zeta)$ is computed by a large and complex computer simulation model whose source code is proprietary and hence unavailable. Obviously apart from making the computation of $F(x, \zeta)$ very expensive, such a situation also precludes the use of automatic differentiation techniques to calculate the gradient $\nabla_x F(x, \zeta)$ (if we know it exists).

If $F(\cdot, \zeta)$ is continuously differentiable on $\mathcal{E}$ for $\mathcal{Q}$-almost all $\zeta$, then it is easy to see that $f$ is also continuously differentiable on $\mathcal{E}$. However, many examples of stochastic optimization problems exist where the $f$ is continuously differentiable on $\mathcal{E}$ even when $F(\cdot, \zeta)$ is only Lipschitz continuous on $\mathcal{E}$ for any $\zeta \in \Xi$. The following is a simple example of such a situation.

Example 1.1. Consider the well known news-vendor problem. A company has to decide the quantity $x$ of a seasonal product, to order. The product can be purchased at a unit cost of $c$ and sold at a unit price of $r > c$ while in season. After that however, the remaining unsold stock can only be sold at a unit salvage price of $s$, where $s < c$. Suppose further, that the demand $\zeta \in \mathbb{R}_+$ for that product is uncertain and has a distribution function $G : \mathbb{R} \rightarrow [0, 1]$ associated with it that is continuous on $\mathbb{R}$. In order to make the
decision regarding the quantity $x$, the company wishes to solve the optimization problem

$$\max_{x \in [0, \infty)} f(x) := \mathbb{E}[F(x, \zeta)]$$

where the profit $F : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}$ can be written for any fixed $x$ and $\zeta$ as

$$F(x, \zeta) := \begin{cases} 
(s - c)x + (r - s)\zeta & \text{if } x \geq \zeta \\
(r - c)x & \text{if } x < \zeta
\end{cases}$$

Indeed, for any given $\zeta \geq 0$, $F(\cdot, \zeta)$ is not differentiable at $x = \zeta$.

However, the expected value of the profit, $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ can be shown to be given for any $x \in \mathbb{R}_+$, by

$$f(x) = \mathbb{E}[F(x, \zeta)] = \int_{(-\infty, \infty)} F(x, \zeta) dG(\zeta) = (r - c)x - (r - s) \int_{[0, x]} G(\zeta) d\zeta$$

Since $G$ is known to be continuous on $\mathbb{R}$, it follows that $f$ is continuously differentiable on $\mathbb{R}_+$.

It can be shown that the same situation occurs in many two stage stochastic linear programming problems, where the random data have densities associated with their distributions.

Assumption A 1 immediately suggests that any sampling-based solution methodology used on $(\mathbb{P})$ should attempt to incorporate the following features.

- The sample sizes used in the algorithm have to remain manageable small. Of course, more precise statements regarding the sample size can be made only if the computational effort required to evaluate $F(x, \zeta)$ and the total available computational budget are known.
- Once $F(x, \zeta)$ is evaluated for a given $x$ and $\zeta$, this value must be reused as much as possible.

Now, let us again consider the two approaches to solving the problem $(\mathbb{P})$ that we discussed earlier, in light of Assumption A 1 and the above requirements. In the case of stochastic approximation, the same algorithmic recursion as in (2) can be used along with a biased gradient estimator which is generated as follows. Given $x_n \in \mathcal{X}$, $c_n > 0$ and some $N \in \mathbb{N}$, we generate independent realizations $\zeta_{n}^{ij} \in \Xi$ for $i = 0, \ldots, l$ (where $l$ is the dimension of the search space) and $j = 1, \ldots, N$, that are also independent of all realizations generated earlier in the procedure. Then, we define the gradient estimator as

$$\nabla \hat{f}(x_n) := \begin{pmatrix}
\sum_{j=1}^{N} F(x_n + c_ne^{j\epsilon}, \zeta_{n}^{ij}) - \sum_{j=1}^{N} F(x_n, \zeta_{n}^{ij}) \\
\vdots \\
\sum_{j=1}^{N} F(x_n + c_ne^{l\epsilon}, \zeta_{n}^{ij}) - \sum_{j=1}^{N} F(x_n, \zeta_{n}^{ij})
\end{pmatrix}_{Nc_n}$$
where $e^i$ denotes the unit vector in the $i$-th coordinate direction in $\mathbb{R}^l$. Essentially, this is a finite difference gradient approximation where the function values $f(x_n + c_ne^i)$ for $i = 1, \ldots, l$ are approximated by sample averages found using independently generated samples. This idea, first suggested in $?$, has the advantage that convergence can be shown even for a sample size $N = 1$. However, it is explicitly required in the algorithm that each new gradient estimate be generated using independent samples. Therefore, each iteration of the algorithm requires $(l + 1) \times N$ new evaluations of the function $F$. Thus, given also the slow progress of this algorithm due to small step-sizes, the stated algorithm may not be suitable for solving $(\mathbb{P})$ under Assumption A 1. However, subsequent research in this field has resulted in the development of gradient approximations that can be generated using at most $2N$ evaluations of $F$. Algorithms that use such gradient approximations are collectively known in literature as Simultaneous Perturbation Stochastic Approximation (SPSA). A combination of these gradient approximations, used along with adaptive step sizes, may prove useful in solving $(\mathbb{P})$. We refer the reader to $?$ for the basic SPSA algorithm and its convergence analysis.

In this paper however, in order to be able to reuse evaluations of $F$, we will not consider algorithms that require independent samples to be generated at each iteration. Accordingly, let us look at the sample average approximation approach in light of Assumption A 1. As we noted earlier, having fixed the sample size $N$ and some $\bar{\omega} \in \Omega$, $(\hat{\mathbb{P}})$ becomes a deterministic optimization problem. With Assumption A 1 however, it is clear that sensitivity measures of the sample average function $\hat{f}_N(\cdot, \bar{\omega})$ like derivatives or subgradients, are also unavailable. Fortunately, there exist several different types of iterative algorithms for the optimization of deterministic functions, without the use of derivative information. The literature in this area of research is quite extensive but not all the different approaches are suited for solving the problem $(\hat{\mathbb{P}})$ under the assumption that function evaluations are expensive. Hence, we will mention only a couple of ideas that are related to the approach that we propose in this paper, and refer the reader to $?$ for an excellent review of the state of the art in derivative free algorithms.

The simplest approach may be to use a direct search algorithm, i.e., an algorithm that proceeds using only direct comparisons of objective function values at different points. Common examples of such algorithms include the simplex reflection algorithm of $?$ and the Parallel Direct Search and Multi-directional Search algorithms proposed respectively in $?$ and $?$. Indeed, if $\hat{f}_N(\cdot, \bar{\omega})$ is not continuously differentiable on $\mathcal{X}$, then there is often no other alternative but to use such direct search algorithms to solve $(\hat{\mathbb{P}})$. On the other hand, if $\hat{f}_N(\cdot, \bar{\omega})$ is smooth, then such algorithms tend to progress slower than algorithms that exploit the smoothness of $\hat{f}_N$.

When $\hat{f}_N(\cdot, \bar{\omega})$ is smooth, then one strategy is to use traditional algorithms for deterministic smooth optimization but with finite difference approximations of the gradient (and if possible, the Hessian). But again, in general, it is unlikely that function evaluations can be reused in such an algorithm. Indeed, it is extremely improbable that at some iteration $n$, given the current solution $x_n$, the function $\hat{f}_N$ has already
been evaluated at \( x_n + c_n e^i \) where \( e^i \) is a unit vector in some coordinate direction. Thus, it is highly likely that such a method will require \((l+1) \times N\) new function evaluations at each iteration, in order to approximate just the gradient.

Recently however, there has been considerable interest in trust region algorithms for derivative free unconstrained deterministic optimization of smooth functions under the assumption that function evaluations are expensive. The idea is that instead of trying to approximate the unavailable higher order derivatives of the objective function, we could construct a polynomial model that approximates the objective function itself in a neighborhood of interest, using interpolation. Let us explore this approach in greater detail since its core ideas are also applicable to the class of algorithms that we propose in this paper. Assume for the moment that the feasible set \( \mathcal{X} = \mathbb{R}^l \) and that \( \hat{f}_N(\cdot, \omega) : \mathbb{R}^l \rightarrow \mathbb{R} \) is sufficiently smooth. At iteration \( n \), given the current iterate \( x_n \), we construct an interpolation set \( \mathcal{Y}_n \) with the appropriate number of points. That is, if we are interested in approximating \( \hat{f}_N \) with a linear model then there are \( l+1 \) parameters to be determined and hence the set \( \mathcal{Y}_n \) must contain \( l+1 \) points including \( x_n \). Similarly, if the model is quadratic, then \( \mathcal{Y}_n \) must contain \( 1 + l + l(l+1)/2 \) points including \( x_n \). Obviously, we try to include as many points in the interpolation set as possible, at which the function \( \hat{f}_N \) has been evaluated already. Then, we find the parameters in the model \( m_n : \mathbb{R}^l \rightarrow \mathbb{R} \) using the interpolation equations

\[
m_n(x^i) = \hat{f}_N(x^i) \quad \text{for} \quad i = 1, \ldots, |\mathcal{Y}_n| \tag{3}
\]

Finally, the model \( m_n \) is optimized in a trust-region \( T_n := \{ x \in \mathbb{R}^l : \| x - x_n \| \leq \Delta_n \} \) which (hopefully) yields a point with a lower objective function value. The points in the interpolation set \( \mathcal{Y}_n \) and the trust region radius \( \Delta_n \) are updated as the algorithm progresses.

This idea has its origins in an algorithm proposed in \(?\), where the author proposed the use of a quadratic interpolation model. Much later, \(?\) suggested an algorithm for constrained optimization where both the objective function and the constraints were modeled using linear interpolation. However, there is some crucial insight regarding such methods in \(?\) that relates the interpolation set \( \mathcal{Y}_n \) to the quality of the model \( m_n \) as an approximation to the objective function within the trust region. Consider a model function that is traditionally used in trust region algorithms.

\[
m_n(x) := \hat{f}_N(x_n, \tilde{\omega}) + \nabla \hat{f}_N(x_n, \tilde{\omega})^T (x - x_n) + \frac{1}{2} (x - x_n)^T H_n (x - x_n) \tag{4}
\]

In (4), either \( H_n \in \mathbb{S}^{l \times l} \) (where \( \mathbb{S}^{l \times l} \) is the space of real symmetric \( l \times l \) matrices) is either set to be \( \nabla^2 f_N(x_n, \tilde{\omega}) \) or obtained using a quasi-Newton update. In any case, if \( f_N(\cdot, \omega) \in C_2(T_n) \) and \( \| \nabla^2 f_N(x, \omega) \|_2 \leq K_n \) for all \( x \in T_n \), we can derive the following bound on \( |f_N(x, \tilde{\omega}) - m_n(x)| \) for \( x \in T_n \) using the Taylor series
expansion of $f_N(\cdot, \bar{\omega})$ at $x_n$.

$$|f_N(x, \bar{\omega}) - m_n(x)| = \left| f_N(x, \bar{\omega}) - f_N(x_n, \bar{\omega}) - \nabla f_N(x_n, \bar{\omega})^T (x - x_n) - \frac{1}{2} (x - x_n)^T H_n (x - x_n) \right|$$

$$= \frac{1}{2} (x - x_n)^T \nabla^2 f_N (x_n + t_n(x - x_n))(x - x_n) - \frac{1}{2} (x - x_n)^T H_n (x - x_n) \right| \quad \text{where} \quad t_n \in [0, 1]$$

$$\leq \left\| \frac{H_n}{2} + K_n \right\| \Delta_n^2$$

(5)

Such a bound relating the accuracy of the model function to the trust region radius is critical to the proper performance of any trust region algorithm.

Now, instead of (4), suppose a quadratic model given by

$$m_n(x) := \hat{f}_N(x_n, \bar{\omega}) + \beta_n^T (x - x_n) + \frac{1}{2} (x - x_n)^T \Lambda_n (x - x_n),$$

(6)

that is obtained by solving the interpolation equations in (3) for the components of $\beta \in \mathbb{R}^l$ and $\Lambda_n \in \mathbb{S}^{l \times l}$, is used in a trust region algorithm. In order for such an algorithm to be successful, the interpolation points in $Y_n$ must be chosen such that the resulting model satisfies a bound similar to that in (5). It is well known that for a finite difference gradient approximation, the error in the approximation decreases to zero as the step-size used to calculate the gradient approximation reduces to zero. Analogously, the interpolation points in $Y_n$ must be chosen to lie in a sufficiently small neighborhood of $x_n$ in order for $\beta_n$ and $H_n$ to be accurate estimates of $\nabla f_N(x_n, \bar{\omega})$ and $\nabla^2 f_N(x_n, \bar{\omega})$ respectively. Further, it can be shown that under certain conditions, as the points in $Y_n$ get progressively closer to $x_n$, $\beta_n$ converges to $\nabla f_N(x_n, \bar{\omega})$ and $H_n$ converges to $\nabla^2 f_N(x_n, \bar{\omega})$. However, in order to obtain a bound similar to that in (5), it turns out the points in $Y_n$ also have to satisfy certain geometric conditions imposed on their positions relative to $x_n$. In particular, the interpolating points cannot lie on any quadratic surface in $\mathbb{R}^l$. If they do, then it can be shown that the equations in (3) have multiple solutions and that there exist solutions that give rise to models $m_n$ that are bad approximations of the objective function within the trust region. In order to avoid such a situation, in practice we usually require that the interpolation points should be sufficiently far away from any quadratic surface in $\mathbb{R}^l$. An interpolation set that satisfies such a condition is referred to as being well poised.

? defined the set $Y_n$ for each $n \in \mathbb{N}$ to consist of $l + l(l + 1)/2$ points chosen entirely from points at which the objective function had been previously evaluated and assumed that at each $n \in \mathbb{N}$, $Y_n$ would automatically satisfy the geometric requirement mentioned above. ? however, ensured that the interpolation set is updated in such a way that it continues to be well poised throughout. All subsequent algorithms that fit in this framework have included in some form or the other, methods to keep the interpolation set well-poised.

? suggested a variant that used quadratic models and showed how Lagrange interpolation polynomials could be used to add new points to the interpolation set to ensure its well-poisedness. The same paper also reported some promising numerical results. Further, ? gave the first convergence proof for such algorithms and also showed how Newton Fundamental Polynomials can be used to build quadratic models even when there are
not sufficiently many points in the interpolation set to exactly specify a fully quadratic model. provides a good introduction to the use of Lagrange and Newton polynomials in order to maintain the well-poisedness of the interpolation set. proposed an interesting variant of this idea, where instead of evaluating the objective function at new points purely in order to keep the interpolation set well-poised, the model function is minimized over a modified (albeit non-convex) trust region, which ensures that the minimizer, when added to the interpolation set, will keep it well-poised. The authors show significant gains over the use of finite difference gradients along with quasi-Newton updates for the Hessian. In a series of recent papers, and has revisited the idea of using Lagrange interpolation polynomials and has also investigated quasi-Newton-like updating formulas for the Hessian matrix of the model found using interpolation, in an attempt to reduce the routine linear algebra work in each iteration.

Indeed, the computational experience reported in the references given above seems to indicate that if the sample average function is smooth enough, then solving the sample average problem using such interpolation models can result in substantial reductions in the number of required evaluations of the function $F$. However, rather than working with a fixed sample size $N$, we believe that further reduction in the number of evaluations of $F$ can be achieved by gradually increasing the sample size in an adaptive manner as the algorithm progresses. The following considerations motivate this belief.

Recall that we are looking to solve the problem and we wish to do this by generating a sequence $\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X}$ of points with successively lower objective function values. However, instead of $f$, we have access only to a sequence of sample average functions $\{\hat{f}_N\}_{N \in \mathbb{N}}$ that converge to $f$ as the sample size $N$ increases to infinity. Thus, by solving the sample average problem, we generate a sequence with successively lower values of the sample average function $\hat{f}_N$ and hope that this sequence also achieves reduction in the objective function. Under this setting, the following rationale naturally emerges.

- In the initial stages of the algorithm, when large reductions are possible in the value of $f$, a crude approximation of the objective function may be sufficient to generate points with lower values of $f$. Accordingly, a sample average approximating function $\hat{f}_N$ obtained using a small sample size $N$ may be sufficient to achieve improvement. Therefore, it makes sense to start the optimization process with a small sample size.

- Further, we should not use a large number of function evaluations in generating minute reductions in the value of the sample average function unless we are confident that this will also lead to improvements in the value of $f$. Therefore, we use a sample size $N$ only as long as the reductions obtained in value of $\hat{f}_N$ are significant compared to some measure of the error between $\hat{f}_N$ and $f$.

With this survey of some of the issues involved in the solution of and various algorithmic approaches, next we turn to a class of algorithms that incorporates the aforementioned refinements.
2 A Class of Derivative Free Trust Region Algorithms

In this paper, we propose and analyze a class of iterative algorithms for solving the problem (P) under Assumption A 1, when \( f \) is continuously differentiable on \( \mathcal{X} \). Our class of algorithms incorporates Monte Carlo sampling techniques along with the use of polynomial model functions in a trust region framework.

When the objective function \( f: \mathcal{X} \rightarrow \mathbb{R} \) and its gradient \( \nabla f(x) \) can be evaluated easily, a typical trust region algorithm used to find the stationary points of \( f \) in \( \mathcal{X} \), works as follows.

**Algorithm 1.** Let us set the constants used in the algorithm as

\[
0 < \eta_1 \leq \eta_2 < 1 \quad 0 < \sigma_1 \leq \sigma_2 < 1 \quad \Delta_{\text{max}} > 0
\]

Also let the initial feasible point be \( x_0 \in \mathcal{X} \) and the initial trust region radius be \( 0 < \Delta_0 < \Delta_{\text{max}} \). Then, for any iteration \( n \) and current solution \( x_n \in \mathcal{X} \), we generate the next point \( x_{n+1} \) in the following manner.

**Step 1:** Define a model function \( m_n: \mathcal{X} \rightarrow \mathbb{R} \) that approximates \( f \) within a trust region \( T_n : = \{ x_n + d : \|d\| \leq \Delta_n \} \).

**Step 2:** Find \( x_n + d_n \in \arg\min \{ m_n(x) : x \in \mathcal{X} \cap T_n \} \).

**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,

\[
\Delta_{n+1} \in \begin{cases} 
[\Delta_n, \Delta_{\text{max}}] & \text{if } \rho_n \geq \eta_2 \\
[\sigma_2 \Delta_n, \Delta_n] & \text{if } \rho_n \in [\eta_1, \eta_2) \\
[\sigma_1 \Delta_n, \sigma_2 \Delta_n] & \text{if } \rho_n < \eta_1 
\end{cases}
\]

Thus, at any iteration \( n \), given the current iterate \( x_n \in \mathcal{X} \), we first define trust region \( T_n \), which is a ball of radius \( \Delta_n \), centered at \( x_n \), and defined in the norm \( \| \cdot \| \) on \( \mathbb{R}^l \). We will refer to \( \Delta_n \) as the *trust region radius* and \( \| \cdot \| \) as the *trust region norm*. Next, we define a model function \( m_n: \mathcal{X} \rightarrow \mathbb{R} \) that approximates the function \( f \) in \( \mathcal{X} \cap T_n \). Since \( f(x_n) \) and \( \nabla f(x_n) \) can be evaluated easily, \( m_n \) is defined to be a quadratic function of the form,

\[
m_n(x) = f(x_n) + \nabla f(x_n)^T (x - x_n) + \frac{1}{2} (x - x_n)^T H_n (x - x_n)
\]  \( (7) \)

where \( H_n \in \mathbb{S}^{l \times l} \). Further, if \( \nabla^2 f(x_n) \) is also available, then we set \( H_n = \nabla^2 f(x_n) \). Otherwise \( H_n \) may be obtained for example, via a quasi-Newton update. After defining the model function and trust region, we find an approximate minimizer \( x_n + d_n \) for \( m_n \) within the trust region \( T_n \), that satisfies a *minimum improvement* condition (which we will later elaborate on). We evaluate the relative improvement \( \rho_n \), which
is the actual decrease in the value of $f$ at $x_n + d_n$ as compared to $x_n$, divided by the decrease predicted by $m_n$. If there is sufficient decrease in the function $f$ then we update the solution for the next iteration as $x_{n+1} = x_n + d_n$ and otherwise we set $x_{n+1} = x_n$. Finally, depending on the relative improvement, we either reduce, leave unchanged, or increase the trust region radius $\Delta_n$ for the next iteration.

In order to solve the optimization problem (P) under Assumption A 1, using a trust region algorithm, several modifications have to be made to Algorithm 1. First, since $f(x)$ cannot be evaluated exactly for any $x \in \mathbb{R}^l$, we will instead have to use sample average approximations where ever required. The use of sampling in our algorithms is akin to the sample average approximation approach. That is, we fix $\bar{\omega} \in \Omega$ and generate the sequence $\{\tilde{\zeta}_i := \zeta_i(\bar{\omega})\}_{i \in \mathbb{N}} \subset \Xi$ before we begin the optimization process. All the sample averages required by the algorithm, are taken with respect to this sample. Therefore, as far as our algorithms are concerned, our sample average functions are approximations to the gradient and Hessian, and use the sample average $\hat{f}(x, \tilde{\zeta}_i)$ to each of the design points and find $\hat{\nabla}_n f(x_n)$ at $x_n$ using sample sizes $N_n \in \mathbb{N}$. Accordingly, it will be convenient for us to discard the dependence on $\omega \in \Omega$ from the notation for the sample average functions and denote the sample average function for any $x \in \mathbb{R}^l$ and a sample size $N \in \mathbb{N}$ as $\hat{f}(x, N)$, where $\hat{f}: \mathbb{R}^l \times \mathbb{N} \rightarrow \mathbb{R}$ is given by

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

(8)

Obviously, the model function defined in (7) cannot be used since neither $f(x_n)$ nor $\nabla f(x_n)$ can be evaluated exactly. Therefore, we propose the following alternative model function.

$$m_n(x) := \hat{f}(x_n, N_n^0) + \hat{\nabla}_n f(x_n)^T (x-x_n) + \frac{1}{2} (x-x_n)^T \hat{\nabla}_n^2 f(x_n)(x-x_n)$$

(9)

In (9), since $f$ cannot be evaluated exactly, we fix a sample size $N_n^0$ and use the sample average $\hat{f}(x_n, N_n^0)$ to approximate $f(x_n)$. Similarly, $\hat{\nabla}_n f(x_n) \in \mathbb{R}^l$ and $\hat{\nabla}_n^2 f(x_n) \in \mathbb{S}^{l \times l}$ are approximations to the gradient $\nabla f(x_n)$ and Hessian $\nabla^2 f(x_n)$ respectively, obtained as follows. We evaluate the sample averages $\hat{f}(x_n + y_n^i, N_n^i)$ at $M_n$ points $\{x_n + y_n^i : i = 1, \ldots, M_n\}$ in a neighborhood of $x_n$ using sample sizes $N_n^i \in \mathbb{N}$. We will refer to the points $\{x_n + y_n^i : i = 1, \ldots, M_n\}$ as design points and refer to the corresponding vectors as $\{y_n^i : i = 1, \ldots, M_n\}$ as perturbations. Then, we assign non-negative weights $\{w_n^i, \ldots, w_n^{M_n}\}$ to each of the design points and find $\hat{\nabla}_n f(x_n)$ and $\hat{\nabla}_n^2 f(x_n)$ such that,

$$\left(\hat{\nabla}_n f(x_n), \hat{\nabla}_n^2 f(x_n)\right) \in \arg \min_{(\beta, A) \in \mathbb{R}^l \times \mathbb{S}^{l \times l}} \sum_{i=1}^{M_n} \left[w_n^i \left(\hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) - \beta y_n^i - \frac{1}{2} y_n^i A y_n^i\right)^2\right]\right)$

(10)

Thus, our model function $m_n$ for each $n \in \mathbb{N}$ is a quadratic polynomial which best fits (in the least squares sense) the sample average function values at the design points. We will refer to $m_n$ as defined in (9) as the regression model function, since we essentially perform linear regression to obtain the model.

It is well known that the success of any trust region algorithm depends crucially on its ability to monitor and adaptively improve the accuracy of the model function in approximating the objective function within
the trust region. When a Taylor series based model function as in (7) is used in Algorithm 1, Steps 3 and 4 serve this purpose. At iteration \( n \), the relative improvement \( \rho_n \) evaluated in Step 3, serves as an indicator for the quality of the model. Whenever \( \rho_n \) is small (\( \rho_n < \eta_1 \)), a reduction in the trust region radius in Step 4, improves the accuracy of the model within the trust region.

However, when a regression model function as in (9) is used, its accuracy in approximating \( f \) within the trust region depends not only on the trust region radius, but also on the quality of \( \hat{f}(x_n, N^0_n) \), \( \nabla_n f(x_n) \) and \( \nabla^2_n f(x_n) \) as approximations \( f(x_n), \nabla f(x_n) \) and \( \nabla^2 f(x_n) \) respectively. Therefore, apart from controlling the trust region radius, we must appropriately choose \( N^0_n \), the number of design points \( M_n \), the location of the design points \( \{x_n + y^i_n : i = 1, \ldots, M_n\} \), the corresponding sample sizes \( \{N^i_n : i = 1, \ldots, M_n\} \) and weights \( \{w^i_n : i = 1, \ldots, M_n\} \) for each \( n \in \mathbb{N} \) so that the resulting model function may be sufficiently accurate. Perhaps more importantly, in light of Assumption A 1, we must also seek to minimize the number of evaluations of the function \( F \) used in the optimization process.

Accordingly, in Section 3, we consider the accuracy of the regression model function \( m_n \) defined in (9) as an approximation of \( f \) and describe procedures to pick the design points, sample sizes and weights required to construct \( m_n \) with a specified accuracy. Subsequently, we describe the working of a typical trust region algorithm that uses such regression model functions and show its convergence in Section ??.

## 3 The Regression Model Function

In this section, we analyze the properties of the regression model function as defined in (9) and develop practical schemes to appropriately pick the design points, sample sizes and weights required for its construction.

Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) where \( \mathcal{X} \subset \mathbb{R}^l \) is assumed to be closed. For each \( n \in \mathbb{N} \), we suppose that \( \hat{f}(x_n, N^0_n) \) is calculated for some sample size \( N^0_n \). Also, we assume that the sample averages \( \{\hat{f}(x_n + y^i_n, N^i_n) : i = 1, \ldots, M_n\} \) are evaluated at a set of design points \( \{x_n + y^i_n : i = 1, \ldots, M_n\} \subset \mathbb{R}^l \) using the sample sizes \( \{N^i_n : i = 1, \ldots, M_n\} \). Finally, we assume that using a set of non-negative weights \( \{w^i_n : i = 1, \ldots, M_n\} \), \( \nabla_n f(x_n) \in \mathbb{R}^l \) and \( \nabla^2_n f(x_n) \in S^l \) are determined to satisfy

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

(11)

**Note:** Actually, without loss of generality we can assume that \( w^i_n > 0 \) for \( i = 1, \ldots, M_n \). This is because setting \( w^i_n = 0 \) for some \( i \in \{1, \ldots, M_n\} \) is equivalent to not including the design point \( x_n + y^i_n \) in the set used to determine \( \nabla_n f(x_n) \) and \( \nabla^2_n f(x_n) \). Therefore, we assume that only the points that had positive weights associated with them, were included in the set of design points to begin with. For the same reason, we also assume without loss of generality that \( \|y^i_n\|_2 > 0 \) for each \( i = 1, \ldots, M_n \) and \( n \in \mathbb{N} \).
Our aim is to first establish when the accuracy of the regression model function approaches that of the “exact” model function defined in (7) as \( n \to \infty \). In particular, we investigate the conditions that can be placed on the choice of the various quantities required to find \( \hat{f}(x_n, N_n^0) \), \( \hat{n} \), and \( \hat{n}^2 f(x_n) \), that are sufficient to ensure \( |\hat{f}(x_n, N_n^0) - f(x_n)| \to 0 \), \( |\hat{n} f(x_n) - \nabla f(x_n)| \to 0 \) and \( |\hat{n}^2 f(x_n) - \nabla^2 f(x_n)| \to 0 \) as \( n \to \infty \). Let us start with a result regarding \( |\hat{f}(x_n, N_n^0) - f(x_n)| \).

**Lemma 3.1.** Suppose the following assumptions hold.

A 2. For any compact set \( D \subset \mathcal{X}, \) \( f \) is continuous on \( D \) and the sequence \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \) converges uniformly to \( f \) on \( D \).

\[
\lim_{N \to \infty} \sup_{x \in D} |\hat{f}(x, N) - f(x)| = 0 \tag{12}
\]

A 3. The sequence \( \{N_n^0\}_{n \in \mathbb{N}} \) of sample sizes satisfies \( N_n^0 \to \infty \) as \( n \to \infty \).

Then, for any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{C} \subset \mathcal{X} \) such that \( \mathcal{C} \) is compact, we get

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

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

\[
\lim_{n \to \infty} \hat{f}(x_n, N_n^0) = f(\tilde{x})
\]

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

\[
|\hat{f}(x_n, N_n^0) - f(x_n)| \leq \sup_{x \in \mathcal{C}} |\hat{f}(x, N_n^0) - f(x)|
\]

Therefore, taking limits as \( n \to \infty \) and noting that \( N_n^0 \to \infty \), we get

\[
\lim_{n \to \infty} |\hat{f}(x_n, N_n^0) - f(x_n)| \leq \lim_{n \to \infty} \sup_{x \in \mathcal{X}} |\hat{f}(x, N_n^0) - f(x)| = 0
\]

Next, from the continuity of \( f \) on \( \mathcal{X} \), we get

\[
\lim_{n \to \infty} |\hat{f}(x_n, N_n^0) - f(\tilde{x})| \leq \lim_{n \to \infty} |\hat{f}(x_n, N_n^0) - f(x_n)| + \lim_{n \to \infty} |f(x_n) - f(\tilde{x})| = 0
\]

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

Next, we consider the accuracy of \( \hat{n} f(x_n) \) and \( \hat{n}^2 f(x_n) \) as approximations of \( \nabla f(x_n) \) and \( \hat{n}^2 f(x_n) \). In particular, we will first consider conditions sufficient to ensure that

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

\[
\lim_{n \to \infty} \|\hat{n}^2 f(x_n) - \nabla^2 f(x_n)\|_2 = 0
\]
Accordingly, let us define the notation required to represent and characterize the set of optimal solutions on the right side of (11). First, let the perturbation matrix $Y_n \in \mathbb{R}^{M_n \times l}$ be defined as

$$Y_n := \begin{pmatrix} (y_1^1)^T \\ (y_1^2)^T \\ \vdots \\ (y_1^{M_n})^T \end{pmatrix}$$

(13)

Let $N_n := \{N_1^1, \ldots, N_{M_n}^1\}$ denote the set of sample sizes used to evaluate the sample averages at the design points. Define $\hat{f}(x, Y_n, N_n) \in \mathbb{R}^{M_n}$ for any $x \in \mathbb{R}^l$ as

$$\hat{f}(x_n, Y_n, N_n) := \begin{pmatrix} \hat{f}(x + y_1^1, N_n^1) - \hat{f}(x, N_n^1) \\ \hat{f}(x + y_2^1, N_n^1) - \hat{f}(x, N_n^1) \\ \vdots \\ \hat{f}(x + y_{M_n}^1, N_n^1) - \hat{f}(x, N_n^1) \end{pmatrix}$$

(14)

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

$$f(x_n, Y_n) := \begin{pmatrix} f(x + y_1^1) - f(x) \\ f(x + y_2^1) - f(x) \\ \vdots \\ f(x + y_{M_n}^1) - f(x) \end{pmatrix}$$

(15)

Next, we develop notation to represent the quadratic form $y_n^T A y_n^T$ in (11). Let $y \in \mathbb{R}^l$ be any vector and $H \in \mathbb{S}^{l \times l}$ be any symmetric matrix given by

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

Then, the quadratic form $y^T H y$ expands as follows.

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

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

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

Therefore,

$$\frac{1}{2} y^T H y = \sum_{j=1}^{l} \sum_{k=j+1}^{l} h_{jk} y_j y_k + \frac{1}{2} \sum_{j=1}^{l} h_{jj} y_j^2$$
We wish to write the right side of the above equation as the scalar product of two appropriately defined vectors. Accordingly, we define the vector \( y^Q \in \mathbb{R}^{l(l+1)/2} \) corresponding to \( y \in \mathbb{R}^l \) as

\[
y^Q := \left( y_1 y_2, y_1 y_3, \ldots, y_1 y_l, y_2 y_1, \ldots, y_l y_1, \frac{1}{\sqrt{2}} y_1^2, \ldots, \frac{1}{\sqrt{2}} y_l^2 \right)^T
\]  

\[
= \left( \{y_j y_k\}_{j,k=1}^{l} \cdot \left\{ \frac{1}{\sqrt{2}} y_j^2 \right\}_{j=1}^{l} \right)^T
\]  

(16)

Note the following useful relationship between the Euclidean norms of \( y \) and \( y^Q \):

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

\[
= \frac{1}{2} \left( \sum_{j=1}^{l} y_j^2 \right)^2
\]

\[
= \frac{1}{2} \|y\|_4^2
\]  

(17)

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

\[
H_v := \left( h_{12}, h_{13}, \ldots, h_{1l}, h_{2l}, \ldots, h_{(l-1)l}, \frac{1}{\sqrt{2}} h_{11}, \frac{1}{\sqrt{2}} h_{22}, \ldots, \frac{1}{\sqrt{2}} h_{ll} \right)^T
\]  

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

(18)

Then it is not hard to see that,

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

(19)

Using the notation in (16), let for each \( i \in \{1, \ldots, M_n\} \),

\[
y_i^Q := \left( (y_i^1)_1 (y_i^1)_2, (y_i^1)_1 (y_i^1)_3, (y_i^1)_2 (y_i^1)_3, \ldots, (y_i^l)_1 (y_i^l)_1, (y_i^l)_2 (y_i^l)_l, \ldots, (y_i^l)_{l-1} (y_i^l)_l, \frac{1}{\sqrt{2}} (y_i^1)^2, \ldots, \frac{1}{\sqrt{2}} (y_i^l)^2 \right)^T
\]

where \((y_i^j)_j\) denotes the \( j \)th component of \( y_i^j \). Also, define the matrix \( Y^Q_n \in \mathbb{R}^{M_n \times (l(l+1)/2)} \) as

\[
Y^Q_n := \begin{pmatrix}
y_1^Q \\
y_2^Q \\ \vdots \\
y_n^Q \\
y_{M_n}^Q 
\end{pmatrix}
\]  

(20)

Let for each \( i = 1, \ldots, M_n \), let \( z_i^T := \left( y_i^T (y_i^Q)^T \right) \). Define the regression matrix \( Z_n \in \mathbb{R}^{M_n \times (l(l+1)/2)} \) as

\[
Z_n := \begin{pmatrix}
z_1^T \\
z_2^T \\ \vdots \\
z_{M_n}^T
\end{pmatrix}
\]  

(21)
Obviously, we have $Z_n = \begin{pmatrix} Y_n & Y_n^2 \end{pmatrix}$. Finally, we let $W_n = \mathrm{diag}(w_n^1, \ldots, w_n^{M_n})$. Now, we can write the set of optimal solutions on the right side of (11) as follows.

$$
\arg\min_{(\beta, \Lambda) \in \mathbb{R}^l \times \mathbb{R}^{l \times l}} \sum_{i=1}^{M_n} \left[ w_n^i \left\{ \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) - \beta^T y_n^i - \frac{1}{2} y_n^i T \Lambda y_n^i \right\} \right]^2 = 
$$

$$
\arg\min_{(\beta, \Lambda) \in \mathbb{R}^l \times \mathbb{R}^{l \times (l+1)/2}} \left\| W_n \right\|^2 \left\{ \hat{f}(x_n, Y_n, N_n) - Z_n \left( \begin{array}{c} \beta \\ \Lambda_v \end{array} \right) \right\} \right\|_2^2 \quad (22)
$$

Consider the optimization problem occurring on the right side of (22).

$$
\min_{(\beta, \Lambda) \in \mathbb{R}^{l \times (l+1)/2}} \left\| W_n \right\|^2 \left\{ \hat{f}(x_n, Y_n, N_n) - Z_n \left( \begin{array}{c} \beta \\ \Lambda_v \end{array} \right) \right\} \right\|_2^2 \quad (23)
$$

It is easy to see that this optimization problem is equivalent to computing the weighted projection of the vector $\hat{f}(x_n, Y_n, N_n) \in \mathbb{R}^{M_n}$ on to the subspace of $\mathbb{R}^{M_n}$ spanned by the columns of $Z_n$. We know that such a projection always exists since a subspace is a closed convex set. Hence the set of optimal solutions of the optimization problem in (23) has to be non-empty. Thus, (22) shows that our method of finding such a projection always exists since a subspace is a closed convex set. Hence the set of optimal solutions (22) shows that our method of finding such a projection always exists since a subspace is a closed convex set.

It is well known that the projection problem in (23) can be solved by solving the so-called normal equations associated with this problem. That is, we have

$$
\arg\min_{(\beta, \Lambda) \in \mathbb{R}^l} \left\| W_n \right\|^2 \left\{ \hat{f}(x_n, Y_n, N_n) - Z_n \left( \begin{array}{c} \beta \\ \Lambda_v \end{array} \right) \right\} \right\|_2^2 = 
$$

$$
\left\{ \left( \begin{array}{c} \beta \\ \Lambda_v \end{array} \right) \in \mathbb{R}^p : \left( Z_n^T W_n Z_n \right) \left( \begin{array}{c} \beta \\ \Lambda_v \end{array} \right) = Z_n^T W_n \hat{f}(x_n, Y_n, N_n) \right\}
$$

Using the notation in (18), we define $\left[ \nabla^2 \hat{f}_n(x) \right]_v \in \mathbb{R}^{l(l+1)/2}$ corresponding to $\nabla^2 \hat{f}_n(x)$ by

$$
\left[ \nabla^2 \hat{f}_n(x) \right]_v = \left( \left\{ \nabla^2 \hat{f}_n(x) \right\}_{j,k=1,...,l, j < k}^T, \left\{ \frac{1}{\sqrt{2}} \nabla^2 \hat{f}_n(x) \right\}_{j=1,...,l} \right)^T \quad (24)
$$

where $(\nabla^2 \hat{f}_n(x))_{j,k}$ denotes the element in the $j$th row and $k$th column of $\nabla^2 \hat{f}_n(x)$. Now, we finally get that choosing $\nabla_n f(x_n) \in \mathbb{R}^l$ and $\nabla^2_n f(x_n) \in \mathbb{R}^{l \times l}$ from (11) is equivalent to choosing the components of $\nabla_n f(x_n)$ and $\left[ \nabla^2_n f(x_n) \right]_v$ from a solution of the set of linear equations given by,

$$
\left( Z_n^T W_n Z_n \right) \left[ \begin{array}{c} \nabla_n f(x_n) \\ \nabla^2_n f(x_n) \end{array} \right] = Z_n^T W_n \hat{f}(x_n, Y_n, N_n) \quad (25)
$$
With this notation, we are ready to state conditions on the design points, sample sizes and weights sufficient to ensure that \( \| \nabla_n f(x_n) - \nabla f(x_n) \|_2 \to 0 \) and \( \| \nabla^2_n f(x_n) - \nabla^2 f(x_n) \|_2 \to 0 \) as \( n \to \infty \). Let us begin by considering some intuitive ideas that motivate our conditions.

Consider for a moment, the finite difference approximation \( \hat{\nabla}_h g(x^*) \in \mathbb{R}^l \) of the gradient \( \nabla g(x^*) \) of some function \( g \in C_1(\mathbb{R}^l) \) at some \( x^* \in \mathbb{R}^l \), defined as

\[
\hat{\nabla}_h g(x^*) := \begin{pmatrix}
g(x^* + he^1) - g(x^*) \\
\vdots \\
g(x^* + he^l) - g(x^*)
\end{pmatrix}
\]

where \( \{e^1, \ldots, e^l\} \) represent unit vectors along the coordinate directions and \( h > 0 \) is the step size. It is easy to see that such a gradient approximation is nothing but the gradient of the linear function

\[
m(x) = g(x^*) + \hat{\nabla}_h g(x^*)^T(x - x^*)
\]

that satisfies

\[
\hat{\nabla}_h g(x^*) \in \arg \min_{\beta \in \mathbb{R}^l} \left\{ \sum_{i=1}^{l} \left[ g(x^* + he^i) - g(x^*) - h\beta^T e^i \right]^2 \right\}
\]

Thus, a linear function \( m(x) \) with the finite difference gradient approximation \( \hat{\nabla}_h g(x_n) \), is the linear function that best fits the function values \( \{g(x + he^i) : i = 1, \ldots, l\} \) at the corresponding design points \{\( x + he^i : i = 1, \ldots, l \}\}. Further, it is well known that

\[
\lim_{h \to 0} \left\| \hat{\nabla}_h g(x^*) - \nabla g(x^*) \right\|_2 = 0
\]

That is, the finite difference approximation get progressively more accurate as the step size \( h \), i.e., the Euclidean distance between \( x^* \) and the design points decreases to zero.

It is intuitive to expect that in the same fashion, the accuracies of \( \hat{\nabla}_n f(x_n) \) and \( \hat{\nabla}^2_n f(x_n) \) determined as in (11) depend on the Euclidean distances \( \{\|y^i_n\|_2 : i = 1, \ldots, M_n\} \) of the design points \( \{x_n + y^i_n : i = 1, \ldots, M_n\} \) from \( x_n \). Indeed, we should expect that in order to ensure \( \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 \to 0 \) and \( \left\| \hat{\nabla}^2_n f(x_n) - \nabla^2 f(x_n) \right\|_2 \to 0 \) 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.

\[
D_n := \{ x \in \mathbb{R}^l : \|x - x_n\|_2 \leq \delta_n \}
\] (26)

We will refer to \( \delta_n > 0 \) as the design region radius for iteration \( n \). Without loss of generality, we will assume that \( \|y^i_n\|_2 \leq \delta_n \) for \( i = 1, \ldots, M'_n \) for some \( M'_n \leq M_n \) and \( \|y^i_n\|_2 > \delta_n \) for \( i = M'_n + 1, \ldots, M_n \). 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 \( \hat{\nabla}_n f(x_n) \) to \( \nabla f(x_n) \) and \( \hat{\nabla}^2_n f(x_n) \) to \( \nabla^2 f(x_n) \), we will ensure that
δₙ → 0 as n → ∞ while at the same time requiring a certain number of inner design points exist for each n ∈ N.

For the notation that we defined earlier with regard the 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

\[ Y_I^n := \begin{pmatrix} y_1^n & \cdots & y_M^n \end{pmatrix}^T \quad \text{and} \quad Y_O^n := \begin{pmatrix} y_{M+1}^n & \cdots & y_M^n \end{pmatrix}^T \]

and hence

\[ Y_n = \begin{pmatrix} Y_I^n \\ Y_O^n \end{pmatrix} \quad \text{(27)} \]

and define the matrices \( Y_I^Q, Y_O^Q, Z_I^n \) and \( Z_O^n \) in the same manner. Analogous to \( N_n \), we let \( N_I^n := \{N_1^n, \ldots, N_M^n\} \) and \( N_O^n := \{N_{M+1}^n, \ldots, N_M^n\} \) denote the set of sample sizes used to evaluate the sample average functions at the inner and outer design points respectively. Then, the vectors \( \tilde{f}(x, Y_I^n, N_I^n) \) and \( \tilde{f}(x, Y_O^n, N_O^n) \) are defined analogous to \( \tilde{f}(x, Y_n, N_n) \). Also, we let \( W_I^n = \text{diag}(w_1^n, \ldots, w_M^n) \) and \( W_O^n = \text{diag}(w_{M+1}^n, \ldots, w_M^n) \).

We noted earlier that as \( n \to \infty \) we will ensure that \( \delta_n \to 0 \) as \( n \to \infty \). This implies that

\[ \lim_{n \to \infty} \max_{i=1, \ldots, M^n} \|y_i^n\|_2 = 0 \]

That is, the inner perturbation vectors corresponding to larger \( n \) will be much smaller in norm that those used earlier in the sequence. Therefore, in order to compare various sets of the design points for different \( n \in \mathbb{N} \), we will scale the corresponding perturbations such that the corresponding design region radii are all equal to 1. Accordingly, we define 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 \( \tilde{z}_i^n \in \mathbb{R}^p \)

\[ \tilde{y}_i^n := \frac{y_i^n}{\delta_n} \quad \text{and} \quad \tilde{z}_i^n := \left( \frac{y_i^n}{\delta_n} \right)^Q \quad \text{(28)} \]

The corresponding scaled perturbation and regression matrices are defined as follows.

\[ \tilde{Y}_n := \begin{pmatrix} (\tilde{y}_1^n)^T \\ \vdots \\ (\tilde{y}_M^n)^T \end{pmatrix} = \frac{Y_n}{\delta_n} \quad \text{(29)} \]

\[ \tilde{Y}_n^Q := \begin{pmatrix} \{y_i^n\}^Q \\ \delta_n^2 \end{pmatrix} = \frac{Y_n^Q}{\delta_n^2} \quad \text{(30)} \]

\[ \tilde{Z}_n := \begin{pmatrix} \tilde{Y}_n \\ \tilde{Y}_n^Q \end{pmatrix} = \tilde{Z}_n D_n \quad \text{(31)} \]
where

\[
D_n = \text{diag} \left( \frac{1}{\delta_n}, \ldots, \frac{1}{\delta_n}, \frac{1}{\delta_n^2}, \ldots, \frac{1}{\delta_n^2} \right)
\]

(32)

The corresponding scaled matrices \( \hat{Y}_n^l, \hat{Y}_n^O, (\hat{Y}_n^l)^Q, (\hat{Y}_n^O)^Q, \hat{Z}_n^l \) and \( \hat{Z}_n^O \) are defined analogously.

Also, note that we use sample averages of the form \( \hat{f}(x_n + y_n^i, N_n^i) \) for \( i = 1, \ldots, M_n \) in the calculation of \( \hat{\nabla}_n f(x_n) \) and \( \hat{\nabla}_n^2 f(x_n) \) for each \( n \in \mathbb{N} \). In order to obtain \( \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 \rightarrow 0 \) and \( \left\| \hat{\nabla}_n^2 f(x_n) - \nabla^2 f(x_n) \right\|_2 \rightarrow 0 \) as \( n \rightarrow \infty \), since \( \nabla f(x_n) \) and \( \nabla^2 f(x_n) \) are quantities related to the function \( f \), it is intuitive to expect that the sequence \( \{ \hat{f}(\cdot, N) \}_{N \in \mathbb{N}} \) must converge to the function \( f \) on \( \mathcal{X} \), in a certain sense as the sample size \( N \) grows to \( \infty \). For example, we already know from the strong law of large numbers that the sequence \( \hat{f}(x, N) - f(x) \rightarrow 0 \) point wise for each \( x \in \mathbb{R}^l \) such that \( f(x) \) is finite for \( \mathcal{P} \)-almost all \( \omega \). However, we require stronger forms of convergence of the sample average function \( \hat{f}(\cdot, N) \) to \( f \). With regard to such a requirement, let us define notation for the required function spaces and their norms. First, for any \( \mathcal{D} \subset \mathbb{R}^l \), let \( \mathcal{W}_0(\mathcal{D}) \) denote the space of all Lipschitz continuous functions \( \phi: \mathcal{D} \rightarrow \mathbb{R} \). The standard norm (referred to as the Lipschitz norm) defined on \( \mathcal{W}_0(\mathcal{D}) \) is as follows. For any \( \phi \in \mathcal{W}_0(\mathcal{D}) \),

\[
\|\phi\|_{\mathcal{W}_0(\mathcal{D})} := \sup_{x \in \mathcal{D}} |\phi(x)| + \sup_{x, x + y \in \mathcal{D}} \frac{|\phi(x) - \phi(y)|}{\|y\|_2} < \infty
\]

As usual, we let \( \mathcal{C}_1(\mathcal{D}) \) denote the space of continuously differentiable functions on \( \mathcal{D} \) with

\[
\|\phi\|_{\mathcal{C}_1(\mathcal{D})} = \sup_{x \in \mathcal{D}} |\phi(x)| + \sup_{x \in \mathcal{D}} \|\nabla \phi(x)\|_2
\]

Further, let \( \mathcal{W}_1(\mathcal{D}) \) denote the space of Lipschitz continuously differentiable functions on \( \mathcal{D} \). That is \( \phi \in \mathcal{W}_1(\mathcal{D}) \) if

\[
\|\phi\|_{\mathcal{W}_1(\mathcal{D})} := \sup_{x \in \mathcal{D}} |\phi(x)| + \sup_{x \in \mathcal{D}} \|\nabla \phi(x)\|_2 + \sup_{x, x + y \in \mathcal{D}} \frac{\|\nabla \phi(x + y) - \nabla \phi(x)\|_2}{\|y\|_2} < \infty
\]

Finally, we let \( \mathcal{C}_2(\mathcal{D}) \) denote the set of twice continuously differentiable functions on \( \mathcal{D} \). From the aforementioned definitions, it is clear that

\[
\mathcal{W}_0(\mathcal{D}) \supset \mathcal{C}_1(\mathcal{D}) \supset \mathcal{W}_1(\mathcal{D}) \supset \mathcal{C}_2(\mathcal{D})
\]

Finally, using the notation we have developed so far, we note that,

\[
Z_n^T W_n Z_n = \begin{pmatrix}
Y_n^T W_n Y_n & Y_n^T W_n Y_n^Q \\
Y_n^Q T W_n Y_n & Y_n^Q T W_n Y_n^Q
\end{pmatrix}
\]

and

\[
Z_n^T W_n \hat{f}(x_n, Y_n, N_n) = \begin{pmatrix}
Y_n^T W_n \hat{f}(x_n, Y_n, N_n) \\
Y_n^Q T W_n \hat{f}(x_n, Y_n, N_n)
\end{pmatrix}
\]

Therefore, the system of equations in (25) can be divided into two sets of equations given below.

\[
(Y_n^T W_n Y_n) \nabla \hat{f}_n(x_n) + (Y_n^T W_n Y_n^Q) \left[ \nabla^2 \hat{f}_n(x_n) \right]_v = Y_n^T W_n \hat{f}(x_n, Y_n, N_n)
\]

(33)

\[
(Y_n^Q T W_n Y_n) \nabla \hat{f}_n(x_n) + (Y_n^Q T W_n Y_n^Q) \left[ \nabla^2 \hat{f}_n(x_n) \right]_v = Y_n^Q T W_n \hat{f}(x_n, Y_n, N_n)
\]

(34)
While considering sufficient conditions to ensure that \( \left\| \nabla_n f(x_n) - \nabla f(x) \right\|_2 \to 0 \) as \( n \to \infty \), we will assume that \( \nabla_n f(x_n) \) and \( \nabla_n^2 f(x_n) \) are picked satisfy only the first of the two sets of equations, i.e., the set given in (33). Obviously, this is a weaker requirement that picking \( \nabla_n f(x_n) \) and \( \nabla_n^2 f(x_n) \) such that both (33) and (34) are satisfied. Therefore the following result holds when \( \nabla_n f(x_n) \) and \( \nabla_n^2 f(x_n) \) are picked such that (25) holds.

**Theorem 3.2.** Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) and let for each \( n \in \mathbb{N} \), \( \nabla_n f(x_n) \in \mathbb{R}^l \) and \( \nabla_n^2 f(x_n) \in \mathbb{S}^{l \times l} \) be picked so as to satisfy (33). Suppose that the following assumptions hold.

**A 4.** For any compact set \( D \subset \mathcal{E} \), we have that \( f \in C_1(D) \), \( \{\tilde{f}(: ,N)\}_{N \in \mathbb{N}} \subset \mathcal{W}_0(D) \) and

\[
\lim_{N \to \infty} \left\| \tilde{f}(:,N) - f \right\|_{\mathcal{W}_0(D)} = 0
\]

**A 5.** The sample sizes \( N^i_n \) corresponding to the inner design points all increase to infinity as \( n \to \infty \). That is,

\[
\lim_{n \to \infty} \min_{i = 1, \ldots, M^i} N^i_n = \infty
\]

**A 6.** 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 \( n \to \infty \).

**A 7.** The set of design points \( \{x_n + y^i_n : i = 1, \ldots, M_n \text{ and } n \in \mathbb{N}\} \) and the sequence \( \{W_n\}_{n \in \mathbb{N}} \) of weight matrices, satisfy the following.

1. There exists a compact set \( C \subset \mathcal{E} \) such that the point \( x_n \) and the set of design points \( \{x_n + y^i_n : i = 1, \ldots, M_n\} \) satisfy

\[
x_n \in C \text{ and } \{x_n + y^i_n : i = 1, \ldots, M_n\} \subset C
\]

for each \( n \in \mathbb{N} \).

2. There exists \( K^j_n \subset \mathbb{R} \) such that for each \( n \in \mathbb{N} \),

\[
\left\| (\tilde{Y}_n)^T W_n \tilde{Y}_n \right\|_2 \left\| (\tilde{Y}_n)^T W_n \tilde{Y}_n \right\|_2^{-1} < K^j_n
\]

3.

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

**A 8.** The sequence of Hessian approximation matrices \( \{\nabla^2 \tilde{f}_n(x_n)\}_{n \in \mathbb{N}} \) is norm-bounded. That is, there exists \( K_H \subset \mathbb{R} \) such that \( \left\| \nabla^2 \tilde{f}_n(x_n) \right\|_2 \leq K_H \) for all \( n \in \mathbb{N} \).

Then, it holds that

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

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

\[
\lim_{n \to \infty} \nabla_n f(x_n) = \nabla f(\bar{x})
\]
We will first show a couple of useful Lemmas and then proceed with the proof of Theorem 3.2. For the next lemma, and for the rest of this article, we will let $\lambda^{\text{max}}(A)$ and $\lambda^{\text{min}}(A)$ denote respectively the maximum and minimum eigenvalues of any square matrix $A$.

**Lemma 3.3.** Let Assumption A 7 hold. Then, $Y_n^T W_n Y_n$ is positive definite for each $n \in \mathbb{N}$ and the following assertions hold.

1. 
\[
\left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\| Y_n^{I^T} W_n^I Y_n^I \right\|_2 < K^I_\lambda \quad \text{for all } n \in \mathbb{N} \tag{38}
\]

2. 
\[
\lim_{n \to \infty} \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\| Y_n^{O^T} W_n^O Y_n^O \right\|_2 = 0 \tag{39}
\]

**Proof.** First, from the definition of $\tilde{Y}_n$, we can see that
\[
\left\| (\tilde{Y}_n^I)^T W_n^I \tilde{Y}_n^I \right\|_2 \left\| ((\tilde{Y}_n^I)^T W_n^I \tilde{Y}_n^I)^{-1} \right\|_2 = \left\| Y_n^{I^T} W_n^I Y_n^I \right\|_2 \left\| Y_n^{I^T} W_n^I Y_n^I \right\|_2 \quad \text{and}
\]
\[
\left\| (\tilde{Y}_n^O)^T W_n^O \tilde{Y}_n^O \right\|_2 \left\| ((\tilde{Y}_n^O)^T W_n^O \tilde{Y}_n^O)^{-1} \right\|_2 = \left\| Y_n^{O^T} W_n^O Y_n^O \right\|_2 \left\| Y_n^{O^T} W_n^O Y_n^O \right\|_2
\]

Therefore, (36) and (37) imply that
\[
\left\| (Y_n^{I^T} W_n^I Y_n^I) \right\|_2 \left\| ((Y_n^{I^T} W_n^I Y_n^I)^{-1}) \right\|_2 < K^I_\lambda \quad \text{for all } n \in \mathbb{N} \tag{40}
\]

and
\[
\lim_{n \to \infty} \frac{\left\| (Y_n^{O^T} W_n^O Y_n^O) \right\|_2}{\left\| (Y_n^{I^T} W_n^I Y_n^I) \right\|_2} = 0 \tag{41}
\]

Now since $Y_n^{I^T} W_n^I Y_n^I$ is symmetric and positive semidefinite, we get from (40) that
\[
\lambda^{\text{min}}(Y_n^{I^T} W_n^I Y_n^I) = \frac{1}{\left\| ((Y_n^{I^T} W_n^I Y_n^I)^{-1}) \right\|_2} > 0
\]

Further, we also have,
\[
Y_n^T W_n Y_n = Y_n^{I^T} W_n^I Y_n^I + Y_n^{O^T} W_n^O Y_n^O
\]

Now, since $Y_n^{O^T} W_n^O Y_n^O$ is positive semidefinite, from the interlocking eigenvalues theorem, we know that
\[
\lambda^{\text{min}}(Y_n^T W_n Y_n) \geq \lambda^{\text{min}}(Y_n^{I^T} W_n^I Y_n^I) > 0
\]

Therefore, $Y_n^T W_n Y_n$ is positive definite and
\[
\left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 = \frac{1}{\lambda^{\text{min}}(Y_n^T W_n Y_n)} \leq \frac{1}{\lambda^{\text{min}}(Y_n^{I^T} W_n^I Y_n^I)} = \left\| (Y_n^{I^T} W_n^I Y_n^I)^{-1} \right\|_2 \tag{42}
\]

Therefore, we see that for any $n \in \mathbb{N}$, using (40) and (42),
\[
\left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\| Y_n^{I^T} W_n^I Y_n^I \right\|_2 \leq \left\| (Y_n^{I^T} W_n^I Y_n^I)^{-1} \right\|_2 \left\| Y_n^{I^T} W_n^I Y_n^I \right\|_2 \leq K^I_\lambda
\]
Thus, we have shown that (38) holds. Similarly,
\[
\| (Y_n^T W_n Y_n)^{-1} \|_2 \leq \| Y_n^T W_n^O Y_n \|_2 \leq \| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n^I Y_n \|_2 \leq K_\lambda \left( \frac{\| Y_n^T W_n^O Y_n \|_2}{\| Y_n^T W_n^I Y_n \|_2} \right)
\]
Using (38) on the right hand side of the above equation, we get
\[
\| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n^O Y_n \|_2 \leq K_\lambda \lim_{n \to \infty} \frac{\| Y_n^T W_n^O Y_n \|_2}{\| Y_n^T W_n^I Y_n \|_2} = 0
\]

Lemma 3.4. 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
\[
a := \begin{pmatrix} \| y_1 \|_2 \alpha_1 \\ \| y_2 \|_2 \alpha_2 \\ \vdots \\ \| y_m \|_2 \alpha_m \end{pmatrix} \quad b := \begin{pmatrix} y_1^T v^1 \\ y_2^T v^2 \\ \vdots \\ y_m^T v^m \end{pmatrix} \quad \text{and} \quad c := \begin{pmatrix} y_{m_1}^T v^1 \\ \vdots \\ y_{m_1}^T v^m \\ \| y_{m_1+1} \|_2 \alpha_{m_1+1} \\ \vdots \\ \| y_m \|_2 \alpha_m \end{pmatrix}
\]
Then,
\[
\| Y^T W a \|_2 \leq \max_{i \in \{1, \ldots, m_1\}} \| y_i \|_2 \alpha_i + \max_{i \in \{m_1+1, \ldots, m\}} \| y_i \|_2 \alpha_i
\]
(43)
\[
\| Y^T W b \|_2 \leq \max_{i \in \{1, \ldots, m_1\}} \| y_i^T \|_2 \| v^i \|_2 + \max_{i \in \{m_1+1, \ldots, m\}} \| y_i \|_2 \alpha_i
\]
(44)
\[
\| Y^T W c \|_2 \leq \max_{i \in \{1, \ldots, m_1\}} \| y_i^T \|_2 \| v^i \|_2 + \max_{i \in \{m_1+1, \ldots, m\}} \| y_i \|_2 \alpha_i
\]
(45)
Proof. We first prove (43).

\[ \|Y^T W a\|_2 = \left\| \sum_{i=1}^{m} y_i w_i \| y_i \|_2 \alpha_i \right\|_2 \]
\[ \leq \sum_{i=1}^{m} \| y_i w_i \| y_i \|_2 \alpha_i \]
\[ = \sum_{i=1}^{m} w_i \| y_i \|_2^2 |\alpha_i| \]
\[ = \sum_{i=1}^{m_1} w_i \| y_i \|_2^2 |\alpha_i| + \sum_{i=m_1+1}^{m} w_i \| y_i \|_2^2 |\alpha_i| \]
\[ \leq \sum_{i=1}^{m_1} w_i \| y_i \|_2^2 \max_{k \in \{1, \ldots, m_1\}} |\alpha_k| + \sum_{i=m_1+1}^{m} w_i \| y_i \|_2^2 \max_{k \in \{m_1+1, \ldots, m\}} |\alpha_k| \]
\[ = \sum_{i=1}^{m_1} \sum_{j=1}^{l} w_i y_{ij}^2 \max_{k \in \{1, \ldots, m_1\}} |\alpha_k| + \sum_{i=m_1+1}^{m} \sum_{j=1}^{l} w_i y_{ij}^2 \max_{k \in \{m_1+1, \ldots, m\}} |\alpha_k| \]
\[ = \text{trace}(Y^T W^T Y^T) \max_{k \in \{1, \ldots, m_1\}} |\alpha_k| + \text{trace}(Y^O^T W^O Y^O) \max_{k \in \{m_1+1, \ldots, m\}} |\alpha_k| \]
\[ \leq l \left[ \|Y^T W^T Y^T\|_{2 \in \{1, \ldots, m_1\}} \max_{\alpha_i} \|y_i\|_2 + \|Y^O^T W^O Y^O\|_{2 \in \{m_1+1, \ldots, m\}} \max_{\alpha_i} \|v_i\|_2 \right] \]

Equation (44) can be established in a similar fashion, by using the Cauchy-Schwartz inequality \( |(y_i)^T v| \leq \|y_i\|_2 \|v\|_2 \).

\[ \|Y^T W b\|_2 \leq \sum_{i=1}^{m} \| y_i w_i(y_i)^T v \|_2 \]
\[ = \sum_{i=1}^{m} w_i \| (y_i)^T v \| \|y_i\|_2 \]
\[ \leq \sum_{i=1}^{m} w_i \| y_i \|_2^2 \|v\|_2 \]
\[ \leq \sum_{i=1}^{m_1} w_i \| y_i \|_2^2 \|v\|_2 + \sum_{i=m_1+1}^{m} w_i \| y_i \|_2^2 \|v\|_2 \]
\[ \leq l \left[ \|Y^T W^T Y^T\|_{2 \in \{1, \ldots, m_1\}} \max_{\alpha_i} \|v\|_2 + \|Y^O^T W^O Y^O\|_{2 \in \{m_1+1, \ldots, m\}} \max_{\alpha_i} \|v\|_2 \right] \]

Similarly, we get

\[ \|Y^T W c\|_2 = \left\| \sum_{i=1}^{m_1} y_i w_i(y_i)^T v \right\|_2 + \sum_{i=m_1+1}^{m} y_i w_i \| y_i \|_2 \alpha_i \]
\[ \leq \sum_{i=1}^{m_1} \|y_i w_i(y_i)^T v\|_2 + \sum_{i=m_1+1}^{m} \|y_i w_i \| y_i \|_2 \alpha_i \|_2 \]
\[ \leq \sum_{i=1}^{m_1} w_i \| y_i \|_2^2 \|v\|_2 \|_2 + \sum_{i=m_1+1}^{m} w_i \| y_i \|_2^2 \|v\|_2 \|_2 \]
\[ \leq \sum_{i=1}^{m_1} w_i \| y_i \|_2^2 \|v\|_2 \|_2 + \sum_{i=m_1+1}^{m} w_i \| y_i \|_2^2 \|v\|_2 \|_2 \]
\[ \leq l \left[ \|Y^T W^T Y^T\|_{2 \in \{1, \ldots, m_1\}} \max_{\alpha_i} \|v\|_2 + \|Y^O^T W^O Y^O\|_{2 \in \{m_1+1, \ldots, m\}} \max_{\alpha_i} \|v\|_2 \right] \]
Proof of Theorem 3.2:

From Lemma 3.3, we know that \( Y_n^T W_n Y_n \) is non-singular for each \( n \in \mathbb{N} \). Therefore, we can rewrite (33) for each \( n \in \mathbb{N} \), as

\[
\nabla \hat{f}_n(x_n) = (Y_n^T W_n Y_n)^{-1} Y_n^T W_n \hat{f}(x_n, Y_n, N_n) - (Y_n^T W_n Y_n)^{-1} (Y_n^T W_n Y_n^Q) \left[ \nabla^2 \hat{f}_n(x_n) \right]_v \tag{46}
\]

We proceed by manipulating the above expression for \( \nabla \hat{f}_n(x_n) \) given and showing that \( \| \nabla \hat{f}_n(x_n) - \nabla f(x_n) \|_2 \to 0 \) as \( n \to \infty \). From (46), we get

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

Adding and subtracting appropriate quantities, we get

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

Recall that we assumed without loss of generality, that \( \| y_i^2 \|_2 > 0 \) for all \( i \in \{1, \ldots, M_n\} \) and \( n \in \mathbb{N} \). Using this, we define

\[
a_n := \left( \hat{f}(x_n, Y_n, N_n) - f(x_n, Y_n) \right) \tag{47}
\]

\[
= \begin{pmatrix}
\| y_1^2 \|_2 \frac{(\hat{f}(x_n + y_1^M, N_n^M) - \hat{f}(x_n, N_n)) - (f(x_n + y_1^M) - f(x_n))}{\| y_1^M \|_2} \\
\| y_2^M \|_2 \frac{(\hat{f}(x_n + y_2^M, N_n^M) - \hat{f}(x_n, N_n)) - (f(x_n + y_2^M) - f(x_n))}{\| y_2^M \|_2}
\end{pmatrix}
\]

Also, let

\[
b_n := f(x_n, Y_n) - Y_n \nabla f(x_n)
\]

\[
= \begin{pmatrix}
\hat{f}(x_n + y_1^M) - f(x_n) - y_1^M \nabla f(x_n) \\
\vdots \\
f(x_n + y_{M_n}^M) - f(x_n) - y_{M_n}^M \nabla f(x_n)
\end{pmatrix}
= \begin{pmatrix}
\| y_1^2 \|_2 \frac{(\hat{f}(x_n + y_1^M) - f(x_n) - y_1^M \nabla f(x_n))}{\| y_1^M \|_2} \\
\| y_2^M \|_2 \frac{(\hat{f}(x_n + y_2^M) - f(x_n) - y_2^M \nabla f(x_n))}{\| y_2^M \|_2}
\end{pmatrix}
\]

Finally, we set

\[
c_n = Y_n^Q \left[ \nabla^2 f(x_n) \right]_v \tag{49}
\]

\[
= \begin{pmatrix}
(y_1^Q)^T \left[ \nabla^2 f(x_n) \right]_v \\
(y_2^M)^T \left[ \nabla^2 f(x_n) \right]_v \\
\vdots \\
(y_{M_n}^Q)^T \left[ \nabla^2 f(x_n) \right]_v
\end{pmatrix}
\]
Using the notation in (19) and the fact that $\|y_n^i\|_2 > 0$ for all $i = 1, \ldots, M_n$ and $n \in \mathbb{N}$, we get

$$c_n = \left( \frac{1}{2} y_n^1 f(x_n) y_n^1, \ldots, \frac{1}{2} y_n^{M_n} f(x_n) y_n^{M_n} \right) = \left( \frac{1}{2} \|y_n^1\|_2 \frac{y_n^1 f(x_n) y_n^1}{\|y_n^1\|_2}, \ldots, \frac{1}{2} \|y_n^{M_n}\|_2 \frac{y_n^{M_n} f(x_n) y_n^{M_n}}{\|y_n^{M_n}\|_2} \right)$$

Then, using the notation defined above, we get

$$\nabla_n f(x_n) - \nabla f(x) = (Y_n^T W_n Y_n)^{-1} Y_n^T W_n a_n + (Y_n^T W_n Y_n)^{-1} Y_n^T W_n b_n - (Y_n^T W_n Y_n)^{-1} Y_n^T W_n c_n$$

Using the triangle inequality,

$$\left\| \nabla_n f(x_n) - \nabla f(x) \right\|_2 \leq \left\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n a_n \right\|_2 + \left\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n b_n \right\|_2 + \left\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n c_n \right\|_2$$

(51)

Let us consider the three terms on the right hand side of (51) in order.

Using (43) in Lemma 3.2, we get

$$\left\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n a_n \right\|_2 \leq \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\| Y_n^T W_n a_n \right\|_2$$

$$\leq l \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\{ \left\| Y_n^T W_n Y_n \right\|_{2} \max_{i \in \{1, \ldots, M_n\}} \left\{ \left( f(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - (f(x_n + y_n^i) - f(x_n)) \right\} \right\}$$

$$+ \left\{ \left\| Y_n^T W_n Y_n \right\|_{2} \max_{i \in \{M_n + 1, \ldots, M_n\}} \left\{ \left( f(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - (f(x_n + y_n^i) - f(x_n)) \right\} \right\}$$

(52)

Now, we consider the two terms of (52) in order. First, from Lemma 3.3, we know that for all $n \in \mathbb{N}$,

$$\left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\| Y_n^T W_n Y_n \right\|_2 \leq K_n$$

From Assumption A 5, we know that

$$\lim_{n \to \infty} \min_{i = 1, \ldots, M_n} N_n^i = \infty$$

Also, we know that $\{x_n\}_{n \in \mathbb{N}} \subset C$ and $x_n + y_n^i \in C$ for each $i = 1, \ldots, M_n$ and $n \in \mathbb{N}$. Now, from Assumption A 4 and the definition of the norm on $\mathcal{W}_0(C)$, we get for any $\varepsilon > 0$ there exists $N_\varepsilon \in \mathbb{N}$ such that for all $N > N_\varepsilon$,

$$\sup_{x+y, x \in C, y \neq 0} \left\{ \frac{\left| \left( f(x+y, N) - \hat{f}(x, N) \right) - (f(x+y) - f(x)) \right|}{\|y\|_2} \right\} \leq \varepsilon$$

Also, from Assumption A 5, we know that there exists $n_\varepsilon \in \mathbb{N}$ such that for all $n > n_\varepsilon$, $\min\{N_n^i : i = 1, \ldots, M_n\} > N_\varepsilon$. Thus, we get combining these two facts, that for any $\varepsilon > 0$ , there exists $n_\varepsilon \in \mathbb{N}$, such that for all $n > n_\varepsilon$,

$$\max_{i \in \{1, \ldots, M_n\}} \left\{ \frac{\left| \left( f(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - (f(x_n + y_n^i) - f(x_n)) \right|}{\|y_n^i\|_2} \right\} \leq \varepsilon$$
Therefore,
\[
\lim_{n \to \infty} \max_{i \in \{1, \ldots, M^*_n\}} \frac{\left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - \left( f(x_n + y_n^i) - f(x_n) \right) \right|}{\|y_n^i\|_2} = 0
\]

Consequently, we get that,
\[
\lim_{n \to \infty} \left\| (Y^T_n W_n Y_n)^{-1} \right\|_2 \left\| Y^T_n W_n^I Y_n^I \right\|_2 \max_{i \in \{1, \ldots, M^*_n\}} \frac{\left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - \left( f(x_n + y_n^i) - f(x_n) \right) \right|}{\|y_n^i\|_2} = 0
\]

Next consider the second term on the right side of (52). First, from Lemma 3.3, we get
\[
\lim_{n \to \infty} \left\| (Y^T_n W_n Y_n)^{-1} \right\|_2 \left\| Y_n^O^T W_n^O Y_n^O \right\|_2 = 0
\]

Now, we note that for each \( n \in \mathbb{N} \),
\[
\max_{i \in \{M^*_n+1, \ldots, M_n\}} \left| \frac{\hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) - \left( f(x_n + y_n^i) - f(x_n) \right)}{\|y_n^i\|_2} \right| \leq \max_{i \in \{M^*_n+1, \ldots, M_n\}} \left\{ \sup_{x_n + y \in C} \frac{|\hat{f}(x + y, N_n^i) - \hat{f}(x, N_n^i)|}{\|y\|_2} \right\} + \max_{i \in \{M^*_n+1, \ldots, M_n\}} \left\{ \sup_{x_n + y \in C, y \neq 0} \frac{|f(x + y) - f(x)|}{\|y\|_2} \right\}
\]

From Assumption A 4, we know that
\[
\lim_{N \to \infty} \left\| \hat{f}(\cdot, N) - f \right\|_{W_0(C)} = 0
\]

Hence, the sequence \( \left\{ \left\| \hat{f}(\cdot, N)_{W_0(C)} \right\| \right\}_{N \in \mathbb{N}} \) is bounded and there exists \( K_f < \infty \), such that \( \|f\|_{W_0(C)} < K_f \) and \( \sup_{N \in \mathbb{N}} \left\| \hat{f}(\cdot, N)_{W_0(C)} \right\| < K_f \) where \( C \subset \mathcal{E} \) is the compact set mentioned in Assumption A 7. Since we also know from Assumption A 7 that \( x_n \in \mathcal{C} \) for each \( n \in \mathbb{N} \) and \( x_n + y_n^i \in \mathcal{C} \) for each \( n \in \mathbb{N} \), we get
\[
\max_{i \in \{M^*_n+1, \ldots, M_n\}} \left\| \hat{f}(\cdot, N_n^i) \right\|_{W_0(C)} < K_f
\]

Consequently, we get that
\[
\|f\|_{W_0(C)} + \max_{i \in \{M^*_n+1, \ldots, M_n\}} \left\| \hat{f}(\cdot, N_n^i) \right\|_{W_0(C)} \leq 2K_f
\]

Hence, using the fact that \( \lim_{n \to \infty} \left\| (Y^T_n W_n Y_n)^{-1} \right\|_2 \left\| Y_n^O^T W_n^O Y_n^O \right\|_2 = 0 \), we finally get
\[
\lim_{n \to \infty} \left\| (Y^T_n W_n Y_n)^{-1} \right\|_2 \left\| Y_n^O^T W_n^O Y_n^O \right\|_2 \max_{i \in \{M^*_n+1, \ldots, M_n\}} \frac{\left| \left( \hat{f}(x_n + y_n^i, N_n^i) - \hat{f}(x_n, N_n^i) \right) - \left( f(x_n + y_n^i) - f(x_n) \right) \right|}{\|y_n^i\|_2} = 0
\]

Thus, we have shown that
\[
\lim_{n \to \infty} \left\| (Y_n^T Y_n)^{-1} Y_n^T a_n \right\|_2 = 0
\]
Thus \( \delta > 0 \) and \( \ldots \). It follows from (55) that given any \( i \in \{1, \ldots, M_1\} \), we know that \( \ldots \). From Assumption A 4, we know that \( \nabla f : C \rightarrow \mathbb{R}^l \) is continuous on \( C \). Since the set \( C \subset E \) (defined in Assumption A 7) is compact, we get that \( \nabla f \) is uniformly continuous on \( C \). That is, for any \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[
\sup_{x, x+y \in C, 0 < \|y\|_2 \leq \delta} \left| \frac{f(x+y) - f(x) - \nabla f(x)^T y}{\|y\|_2} \right| < \varepsilon
\]  

From Assumption A 7, we know that \( \{x_n\}_{n \in \mathbb{N}} \subset C \) and \( \{x_n + y_n^i : i = 1, \ldots, M_n\} \subset C \) for each \( n \in \mathbb{N} \). Further, we know from Assumption A 6 that \( \delta_n \rightarrow 0 \) as \( n \rightarrow \infty \). 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

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

It follows from (55) that given any \( \delta > 0 \), there exists an \( n_1 \in \mathbb{N} \) such that \( \|x_n + y_n^i - x_n\|_2 < \delta \) for all \( i \in \{1, \ldots, M_n^I\} \) and \( n > n_1 \). Therefore, using (54), we get that for any \( \varepsilon > 0 \), there exists \( n_1 \in \mathbb{N} \), such that

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

Thus

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

Now, using (38) in Lemma 3.3, we get

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

Hence, the first term on the right side of (53) converges to zero. Next we show that

\[
\lim_{n \rightarrow \infty} l \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\| Y_n^O T W_n^O Y_n^O \right\|_2 \max_{i \in \{M_n^O + 1, \ldots, M_n\}} \left| \frac{f(x_n + y_n^i) - f(x_n) - y_n^i T \nabla f(x_n)}{\|y_n^i\|_2} \right| = 0
\]
Again, since $f$ is continuously differentiable on $C$ and $C \subset \mathcal{E}$ is compact, we get from Lemma 2.2 in ?, that there exists $K_{1f} < \infty$ such that

$$
\sup_{x+y, x \in C \atop y \neq 0} \frac{|f(x+y) - f(x) - y^T \nabla f(x)|}{\|y\|_2} < K_{1f}
$$

Since we know from Assumption A 7 that $x_n \in C$ for each $n \in \mathbb{N}$ and $x_n + y_n^i \in C$ for each $n \in \mathbb{N}$ and $i \in \{1, \ldots, M_n\}$, we get that for each $n \in \mathbb{N}$

$$
\max_{i \in \{M_n^i + 1, \ldots, M_n\}} \frac{|f(x_n + y_n^i) - f(x_n) - y_n^i T \nabla f(x_n)|}{\|y_n^i\|_2} \leq K_{1f}
$$

Now, using (39) in Lemma 3.3, we get

$$
\lim_{n \to \infty} \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \|Y_n^{O^T} W_n^{O^O} Y_n^{O^O}\|_{\max_{i \in \{M_n^i + 1, \ldots, M_n\}}} \left\| \frac{f(x_n + y_n^i) - f(x_n) - y_n^i T \nabla f(x_n)}{\|y_n^i\|_2} \right\|
$$

$$
\leq 2l K_{1f} \lim_{n \to \infty} \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\|Y_n^{O^T} W_n^{O^O} Y_n^{O^O}\right\| = 0
$$

Thus, the second term on the right side of (53) converges to zero. Hence $\left\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n b_n \right\|_2 \to 0$ as $n \to \infty$.

Finally, consider the third term on the right in (51). Again using (43),

$$
\left\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n c_n \right\|_2 \leq \frac{l}{2} \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\{ \left\|Y_n^{O^T} W_n^{O^T} Y_n^{O^T}\right\|_{\max_{i \in \{1, \ldots, M_n^i\}}} \left\| \frac{y_n^i T \nabla^2 \hat{f}_n(x_n) y_n^i}{\|y_n^i\|_2} \right\| \right\}
$$

Since $\nabla^2 \hat{f}_n(x_n) \in \mathbb{S}^{l \times l}$ is a symmetric matrix, it is well known that

$$
\left\| \frac{y_n^i T \nabla^2 \hat{f}_n(x_n) y_n^i}{\|y_n^i\|_2} \right\| \leq \left\| \nabla^2 \hat{f}_n(x_n) \right\|_2 \left\| y_n^i \right\|_2
$$

for each $n \in \mathbb{N}$ and $i \in \{1, \ldots, M_n\}$. Using $\left\| y_n^i \right\|_2 > 0$ for each $i = 1, \ldots, M_n$ and $n \in \mathbb{N}$, we get

$$
\left\| \frac{y_n^i T \nabla^2 \hat{f}_n(x_n) y_n^i}{\|y_n^i\|_2} \right\| \leq \left\| \nabla^2 \hat{f}_n(x_n) \right\|_2 \left\| y_n^i \right\|_2 = \left\| \nabla^2 \hat{f}_n(x_n) \right\|_2 \left\| y_n^i \right\|_2
$$

Now, from Assumption A 8 we know $\left\| \nabla^2 \hat{f}_n(x_n) \right\|_2 \leq K_H$ for all $n \in \mathbb{N}$. Hence,

$$
\left\| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n c_n \right\|_2 \leq \frac{lK_H}{2} \left\| (Y_n^T W_n Y_n)^{-1} \right\|_2 \left\{ \left\|Y_n^{O^T} W_n^{O^T} Y_n^{O^T}\right\|_{\max_{i \in \{1, \ldots, M_n^i\}}} \left\| y_n^i \right\|_2 \right\}
$$

$$
+ \left\{ \left\|Y_n^{O^T} W_n^{O^O} Y_n^{O^O}\right\|_{\max_{i \in \{M_n^i + 1, \ldots, M_n\}}} \left\| y_n^i \right\|_2 \right\}
$$

(56)
Corollary 3.6. But we know from Assumption A 6, that \( \max_{i \in \{1, \ldots, M_n^f\}} \|y_n^i\|_2 \to 0 \) as \( n \to \infty \). Therefore, using (38) in Lemma 3.3, we get
\[
\lim_{n \to \infty} \left( \frac{lK_H}{2} \right) \| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n Y_n^f \|_2 \max_{i \in \{1, \ldots, M_n^f\}} \|y_n^i\|_2 \leq \left( \frac{lK_HC}{2} \right) \lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^f\}} \|y_n^i\|_2 = 0
\]

Now, we know from Assumption A 7, that \( \{x_n\}_{n \in \mathbb{N}} \subset C \) and \( x_n + y_n^i \in C \) for all \( i \in \{M_n^f + 1, \ldots, M_n\} \) and \( n \in \mathbb{N} \), where \( C \subset E \) is compact. Therefore, there must exist \( K_y < \infty \) such that \( \max_{i \in \{M_n^f + 1, \ldots, M_n\}} \|y_n^i\|_2 < K_y \) for all \( n \in \mathbb{N} \). Thus, using (39) in Lemma 3.3 we get
\[
\lim_{n \to \infty} \left( \frac{lK_H}{2} \right) \| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n Y_n^f \|_2 \max_{i \in \{1, \ldots, M_n^f\}} \|y_n^i\|_2 \leq \left( \frac{lK_HC}{2} \right) \lim_{n \to \infty} \| (Y_n^T W_n Y_n)^{-1} \|_2 \| Y_n^T W_n Y_n^f \|_2 = 0
\]

Therefore \( \| (Y_n^T W_n Y_n)^{-1} Y_n^T W_n c_n \|_2 \to \infty \) as \( n \to \infty \). Hence we have shown that,
\[
\lim_{n \to \infty} \| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \|_2 = 0
\]

In particular, if \( x_n \to \hat{x} \in \mathcal{X} \) as \( n \to \infty \), then we get from the continuous differentiability of \( f \) that
\[
\lim_{n \to \infty} \| \hat{\nabla}_n f(x_n) - \nabla f(\hat{x}) \|_2 \leq \lim_{n \to \infty} \| \nabla f(x_n) - \nabla f(x_n) \|_2 \leq \lim_{n \to \infty} \| \nabla f(x_n) - \nabla f(\hat{x}) \|_2 = 0
\]

\[\blacksquare\]

**Corollary 3.5.** Consider any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) and let all the assumptions of Theorem 3.2 hold. Suppose for each \( n \in \mathbb{N} \), \( \hat{\nabla}_n f(x_n) \in \mathbb{R}^l \) and \( \hat{\nabla}_n^2 f(x_n) \in S^{l \times l} \) are picked to satisfy (25). Then,
\[
\lim_{n \to \infty} \| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \|_2 = 0
\]

**Proof.** The result follows from the fact that if \( \hat{\nabla}_n f(x_n) \) and \( \hat{\nabla}_n^2 f(x_n) \) satisfy (25) for each \( n \in \mathbb{N} \), then they also automatically satisfy (33). \( \blacksquare \)

**Corollary 3.6.** Consider any sequence \( \{x_n\}_{n \in \mathbb{N}} \subset \mathcal{X} \) and let all the assumptions of Theorem 3.2 hold. Suppose for each \( n \in \mathbb{N} \), \( \hat{\nabla}_n f(x_n) \in \mathbb{R}^l \) is picked such that
\[
(Y_n^T W_n) Y_n \hat{\nabla}_n f(x_n) = Y_n^TW_n f(x_n, Y_n, \mathcal{N}_n)
\]

Then we have,
\[
\lim_{n \to \infty} \| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \|_2 = 0
\]

**Proof.** The required result follows from Theorem 3.2 when we set \( \hat{\nabla}_n^2 f(x_n) = 0 \) in (33), for each \( n \in \mathbb{N} \). \( \blacksquare \)
Next, we present some stronger conditions that ensure the convergence of both the gradient and Hessian approximations.

**Theorem 3.7.** Consider a sequence \( \{x_n\}_{n \in \mathbb{N}} \subset X \) and let for each \( n \in \mathbb{N} \), \( \hat{\nabla}_n f(x_n) \in \mathbb{R}^l \) and \( \hat{\nabla}^2_n f(x_n) \in S^{l \times l} \) be picked so as to satisfy (25). Suppose that the following assumptions hold.

**A 9.** For any compact set \( D \subset E \), we get \( f \in C^2(D) \), \( \{\hat{f}(\cdot, n)\}_{N \in \mathbb{N}} \subset W_1(D) \) and 
\[
\lim_{n \to \infty} \left\| \hat{f}(\cdot, n) - f \right\|_{W_1(D)} = 0
\]

**A 10.** There exists a sequence \( \{N_n^\#\}_{n \in \mathbb{N}} \) of sample sizes such that

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

**A 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 a compact set \( C \subset E \) such that the point \( x_n \) and the set of design points \( \{x_n + y_n^i : i = 1, \ldots, M_n \} \) satisfy
\[
x_n \in C \quad \text{and} \quad \{x_n + y_n^i : i = 1, \ldots, M_n \} \subset C
\]
   for each \( n \in \mathbb{N} \).
2. There exists \( K^I_L < \infty \) such that for each \( n \in \mathbb{N} \)
\[
\left\| (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I \right\|_2 \left\| ((\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I)^{-1} \right\|_2 < K^I_L
\]
3. 
\[
\lim_{n \to \infty} \left\| (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \right\|_2 = 0
\]

Further, let Assumption A 6 also hold. Then, we have,
\[
\lim_{n \to \infty} \left\| \hat{\nabla}_n f(x_n) - \nabla f(x_n) \right\|_2 = 0 \quad \text{and} \quad \lim_{n \to \infty} \left\| \hat{\nabla}^2_n f(x_n) - \nabla^2 f(x_n) \right\|_2 = 0
\]

In particular, if \( \{x_n\}_{n \in \mathbb{N}} \subset X \) is such that \( x_n \to \tilde{x} \in X \) as \( n \to \infty \), then
\[
\lim_{n \to \infty} \hat{\nabla}_n f(x_n) = \nabla f(\tilde{x}) \quad \text{and} \quad \lim_{n \to \infty} \hat{\nabla}^2_n f(x_n) = \nabla^2 f(\tilde{x})
\]

Before proving Theorem 3.7, we state and prove a useful lemma.

**Lemma 3.8.** Let Assumption A 11 hold. Then, \( \tilde{Z}_n^T W_n \tilde{Z}_n \) and \( Z_n^T W_n Z_n \) are positive definite for each \( n \in \mathbb{N} \) and the following hold true.
1. \[ \| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \|_2 \| (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I \|_2 < K_\lambda^I \quad \text{for all} \quad n \in \mathbb{N} \quad (60) \]

2. \[ \lim_{n \to \infty} \| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \|_2 \| (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \|_2 = 0 \quad (61) \]

Proof. Since \((\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I\) is a symmetric and positive semidefinite matrix, we get from (58) that

\[ \lambda^{\min}((\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I) = \frac{1}{\| (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I \|_2} > 0 \]

Further, it is easily seen that

\[ \tilde{Z}_n^T W_n \tilde{Z}_n = (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I + (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \]

Now, since \((\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O\) is positive semidefinite, from the interlocking eigenvalues theorem, we know that

\[ \lambda^{\min}(\tilde{Z}_n^T W_n \tilde{Z}_n) \geq \lambda^{\min}((\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I) > 0 \]

Therefore, \(\tilde{Z}_n^T W_n \tilde{Z}_n\) is positive definite. Further, from the definition of \(D_n\) in (32), 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 \(Z_n^T W_n Z_n = D_n^{-1}(\tilde{Z}_n^T W_n \tilde{Z}_n)D_n^{-1}\) is positive definite for each \(n \in \mathbb{N}\). Now,

\[ \| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \|_2 = \frac{1}{\lambda^{\min}(\tilde{Z}_n^T W_n \tilde{Z}_n)} \leq \frac{1}{\lambda^{\min}((\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I)} = \| (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I \|_2 \quad (62) \]

Therefore, we see that for any \(n \in \mathbb{N}\), using (62),

\[ \| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \|_2 \| (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I \|_2 \leq \| (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I \|_2 \| (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I \|_2 \leq K_\lambda^I \]

Thus, we have shown that (60) holds. Similarly,

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

Using (60) on the right hand side of the above equation, we get

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

Finally, we use (37) in Assumption A 11 to get that

\[ \lim_{n \to \infty} \| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \|_2 \| (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \|_2 \leq K_\lambda^I \lim_{n \to \infty} \| (\tilde{Z}_n^O)^T W_n^O \tilde{Z}_n^O \|_2 \| (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I \|_2 = 0 \]
Proof of Theorem 3.7:

From Lemma 3.8, we know that \( Z_n^TW_nZ_n \) is positive definite and hence non-singular for each \( n \in \mathbb{N} \). Therefore, we can rewrite (25) as

\[
\begin{pmatrix}
\nabla f(x_n) \\
\nabla f(x_n) \\
\n\frac{\partial}{\partial f(x_n)} \\
\n\frac{\partial}{\partial f(x_n)} \\
\end{pmatrix}
\]

(63)

Noting from Assumption A 9 that \( \hat{f}(\cdot, N) \in \mathcal{W}(C) \) for each \( N \in \mathbb{N} \) and from Assumption 10 that \( N_i^0 = N_n^\# \) for \( i = 1, \ldots, M_n \) and since \( \{x_n\}_{n \in \mathbb{N}} \subset C \), we define the vector \( b_n \in \mathbb{R}^{M_n} \) as,

\[
b_n := \begin{pmatrix}
y_1^T(\nabla \hat{f}(x_n, N_n^0) - \nabla f(x_n)) \\
\vdots \\
y_{M_n}^T(\nabla \hat{f}(x_n, N_{M_n}^0) - \nabla f(x_n)) \\
\end{pmatrix} = \left( Y_n(\nabla \hat{f}(x_n, N_n^\#) - \nabla f(x_n)) \right)
\]

Adding and subtracting \( (Z_n^TW_nZ_n)^{-1}Z_n^TW_n(b_n + f(x_n, Y_n)) \) in (63), we get

\[
\begin{pmatrix}
\nabla f(x_n) \\
\nabla f(x_n) \\
\frac{\partial}{\partial f(x_n)} \\
\frac{\partial}{\partial f(x_n)} \\
\end{pmatrix}
\]

(64)

Using the fact that \( (Z_n^TW_nZ_n)^{-1}Z_n^TW_n = D_n(\tilde{Z}_n^TW_n\tilde{Z}_n)^{-1}\tilde{Z}_n^TW_n \), we get

\[
\begin{pmatrix}
\nabla f(x_n) \\
\nabla f(x_n) \\
\frac{\partial}{\partial f(x_n)} \\
\frac{\partial}{\partial f(x_n)} \\
\end{pmatrix}
\]

(65)

We assumed without loss of generality that \( \|y_n^i\|_2 > 0 \) for all \( i \in \{1, \ldots, M_n\} \) and \( n \in \mathbb{N} \). Using (17), it is easily seen that this means \( \|\hat{z}_n^i\|_2 > 0 \) for all \( i \in \{1, \ldots, M_n\} \) and \( n \in \mathbb{N} \). Therefore, again using Assumption A 10, we set for each \( n \in \mathbb{N} \),

\[
a_n := \begin{pmatrix}
\hat{f}(x_n, Y_n, \hat{N}_n) - f(x_n, Y_n) - b_n \\
\hat{f}(x_n, Y_n, \hat{N}_n) - f(x_n, Y_n) - b_n \\
\vdots \\
\hat{f}(x_n, Y_n, \hat{N}_n) - f(x_n, Y_n) - b_n \\
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\|\hat{z}_n^i\|_2 \left( \frac{\hat{f}(x_n + y_n^i, \hat{N}_n^0) - f(x_n, \hat{N}_n^0)}{\|y_n^i\|_2} - (f(x_n + y_n^i) - f(x_n)) - y_n^i^T(\nabla \hat{f}(x_n, \hat{N}_n^0) - \nabla f(x_n)) \right) \\
\|\hat{z}_n^i\|_2 \left( \frac{\hat{f}(x_n + y_n^i, \hat{N}_n^0) - f(x_n, \hat{N}_n^0)}{\|y_n^i\|_2} - (f(x_n + y_n^i) - f(x_n)) - y_n^i^T(\nabla \hat{f}(x_n, \hat{N}_n^0) - \nabla f(x_n)) \right) \\
\vdots \\
\|\hat{z}_n^i\|_2 \left( \frac{\hat{f}(x_n + y_n^i, \hat{N}_n^0) - f(x_n, \hat{N}_n^0)}{\|y_n^i\|_2} - (f(x_n + y_n^i) - f(x_n)) - y_n^i^T(\nabla \hat{f}(x_n, \hat{N}_n^0) - \nabla f(x_n)) \right) \\
\|\hat{z}_n^i\|_2 \left( \frac{\hat{f}(x_n + y_n^i, \hat{N}_n^0) - f(x_n, \hat{N}_n^0)}{\|y_n^i\|_2} - (f(x_n + y_n^i) - f(x_n)) - y_n^i^T(\nabla \hat{f}(x_n, \hat{N}_n^0) - \nabla f(x_n)) \right) \\
\|\hat{z}_n^i\|_2 \left( \frac{\hat{f}(x_n + y_n^i, \hat{N}_n^0) - f(x_n, \hat{N}_n^0)}{\|y_n^i\|_2} - (f(x_n + y_n^i) - f(x_n)) - y_n^i^T(\nabla \hat{f}(x_n, \hat{N}_n^0) - \nabla f(x_n)) \right) \\
\end{pmatrix}
\]

\]

32
Let us simplify the third term on the right side of (65).

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

\[ = \begin{pmatrix} f(x_n + y_n^1) - f(x_n) - y_n^1 T \nabla f(x_n) - (y_n^1)^T \left[ \nabla^2 f(x_n) \right]_v \\ : \\ f(x_n + y_n^M_n) - f(x_n) - y_n^M_n T \nabla f(x_n) - (y_n^M_n)^T \left[ \nabla^2 f(x_n) \right]_v \end{pmatrix} \]

\[ = \begin{pmatrix} f(x_n + y_n^1) - f(x_n) - y_n^1 T \nabla f(x_n) - \frac{1}{2} (y_n^1 T \nabla^2 f(x_n) y_n^1) \\ : \\ f(x_n + y_n^M_n) - f(x_n) - y_n^M_n T \nabla f(x_n) - \frac{1}{2} (y_n^M_n T \nabla^2 f(x_n) y_n^M_n) \end{pmatrix} \]

Again noting that \( \delta_n > 0 \) 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

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

\[ = \begin{pmatrix} \| z_1^n \|_2 \| f(x_n + y_n^1) - f(x_n) - y_n^1 T \nabla f(x_n) - \frac{1}{2} (y_n^1 T \nabla^2 f(x_n) y_n^1) \|_2 \\ : \\ \| z_{M_n}^n \|_2 \| f(x_n + y_n^M_n) - f(x_n) - y_n^M_n T \nabla f(x_n) - \frac{1}{2} (y_n^M_n T \nabla^2 f(x_n) y_n^M_n) \|_2 \end{pmatrix} \]

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

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

\[ \Rightarrow \left\| \left( \frac{\nabla f(x_n)}{[\nabla^2 f(x_n)]_v} \right) - \left( \frac{\nabla f(x_n)}{[\nabla^2 f(x_n)]_v} \right) \right\|_2 \leq \left\| D_n(\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n a_n \right\|_2 + \left\| D_n(\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n b_n \right\|_2 + \left\| D_n(\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n c_n \right\|_2 \] (66)

Next, we show that all three terms on the right side of (66) converge to zero as \( n \to \infty \). But first, we note that from Assumption A 6, \( \delta_n \to 0 \) as \( n \to \infty \). Therefore, there exists \( n^* \in \mathbb{N} \), such that for all \( n > n^* \), \( \delta_n < 1 \). This means that for all \( n > n^* \), \((1/\delta_n)^2 > (1/\delta_n)^2 \). Therefore, we get that for all \( n > n^* \), \( D_n \|_2 = (1/\delta_n)^2 \).

Let us consider the first term right side of (66) for \( n > n^* \). First, we note that

\[ \left\| D_n(\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \hat{Z}_n^T W_n a_n \right\|_2 \leq D_n \left\| (\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \right\|_2 \left\| \hat{Z}_n^T W_n a_n \right\|_2 = \left( 1/\delta_n^2 \right) \left( (\hat{Z}_n^T W_n \hat{Z}_n)^{-1} \right) \| \hat{Z}_n^T W_n a_n \|_2 \]

33
Recall that the matrix $\tilde{Z}_n$ has $p = l + (l + 1)/2$ columns. Using this fact in (43) of Lemma 3.4, we get

$$\left\| D_n(\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \tilde{Z}_n^T W_n a_n \right\|_2 \leq p \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \times \max_{i \in \{1, \ldots, M_n\}} \left\{ \frac{\left\| \hat{f}(x_n + y_n^i, N_n^\#) - \hat{f}(x_n, N_n^\#) \right\|}{\delta_n^2 \left\| z_n^i \right\|_2} - \frac{\left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \times \max_{i \in \{1, \ldots, M_n\}} \left\{ \frac{\left\| \hat{f}(x_n + y_n^i, N_n^\#) - \hat{f}(x_n, N_n^\#) \right\|}{\delta_n^2 \left\| z_n^i \right\|_2} \right\}$$

From the above, we get

$$\left\| D_n(\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \tilde{Z}_n^T W_n a_n \right\|_2 \leq p \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \times \max_{i \in \{1, \ldots, M_n\}} \left\{ \frac{\left\| \hat{f}(x_n + y_n^i, N_n^\#) - \hat{f}(x_n, N_n^\#) \right\|}{\delta_n^2 \left\| z_n^i \right\|_2} \right\}$$

First, from Assumption A 11 and Lemma 3.8 we get that

$$\left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \leq K^4 \quad \text{for all} \quad n \in \mathbb{N}$$

and

$$\lim_{n \to \infty} \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 = 0$$

Therefore, clearly there exists a constant $K_{\lambda} < \infty$ such that for all $n \in \mathbb{N}$,

$$\left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 < K_{\lambda} \quad (67)$$

Next, we note that from (17), for any $n \in \mathbb{N}$ and $i \in \{1, \ldots, M_n\}$,

$$\delta_n^2 \left\| z_n^i \right\|_2 = \sqrt{\delta_n^2 \left\| y_n^i \right\|_2^2 + \left(\frac{1}{2}\right) \left\| y_n^i \right\|_2^4} \geq \left(\frac{1}{2}\right) \left\| y_n^i \right\|_2^2 \quad (68)$$

Therefore, for each $i = 1, \ldots, M_n$,

$$\left\| \hat{f}(x_n + y_n^i, N_n^\#) - \hat{f}(x_n, N_n^\#) \right\| \leq \frac{\left\| \hat{f}(x_n + y_n^i, N_n^\#) - \hat{f}(x_n, N_n^\#) \right\|}{\delta_n^2 \left\| z_n^i \right\|_2} \left\| y_n^i \right\|_2^2$$

Now from Assumption A 10, we know that, $N_n^\# \to \infty$ as $n \to \infty$. From Assumption A 9 we then get that

$$\lim_{n \to \infty} \left\| \hat{f}(\cdot, N_n^\#) - f \right\|_{\mathcal{W}_1(C)} = 0$$
From Assumption A 9 and the fact that $C \subset E$ is the compact set defined in Assumption A 11. Therefore, we get from Lemma 2.4 in ? that
\[
\lim_{n \to \infty} \sup_{y \neq 0} \left\| \left( \hat{f}(x + y, N_n^\#) - f(x + y) \right) - \left( \hat{f}(x, N_n^\#) - f(x) \right) - y^T (\nabla \hat{f}(x, N_n^\#) - \nabla f(x)) \right\|_{2} = 0
\]

Further, from Assumption A 11, we know that $x_n + y_n^i \in C$ for each $i = 1, \ldots, M_n$ and $n \in \mathbb{N}$ and $\{x_n\}_{n \in \mathbb{N}} \subset C$.

Therefore, for each $n \in \mathbb{N},$
\[
\max_{i \in \{1, \ldots, M_n\}} \left\| \left( \hat{f}(x_n + y_n^i, N_n^\#) - \hat{f}(x_n, N_n^\#) \right) - \left( f(x_n + y_n^i) - f(x_n) \right) - y_n^i \left( \nabla \hat{f}(x_n, N_n^\#) - \nabla f(x_n) \right) \right\|_{2} \leq \sup_{x, x + y \in C \atop y \neq 0} \left\| \left( \hat{f}(x + y, N_n^\#) - f(x + y) \right) - \left( \hat{f}(x, N_n^\#) - f(x) \right) - y^T (\nabla \hat{f}(x, N_n^\#) - \nabla f(x)) \right\|_{2}
\]

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

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

Next, let us consider the second term on the right side of (66). Letting $0_{I(l+1)/2}$ denote a column vector with $I(l + 1)/2$ elements each of which is equal to zero, it is easily seen that
\[
b_n = Y_n (\nabla \hat{f}(x_n, N_n^\#) - \nabla f(x_n)) = Z_n \left( \nabla \hat{f}(x_n, N_n^\#) - \nabla f(x_n) \right) = \hat{Z}_n D_n^{-1} \left( \nabla \hat{f}(x_n, N_n^\#) - \nabla f(x_n) \right) 0_{I(l+1)/2}
\]
Using the definition of $b_n$, we get for each $n \in \mathbb{N}$
\[
\left\| D_n \left( \hat{Z}_n^T W_n \hat{Z}_n \right)^{-1} \hat{Z}_n^T W_n b_n \right\|_{2} = \left\| D_n \left( \hat{Z}_n^T W_n \hat{Z}_n \right)^{-1} \hat{Z}_n^T W_n \hat{Z}_n D_n^{-1} \left( \nabla \hat{f}(x_n, N_n^\#) - \nabla f(x_n) \right) 0_{I(l+1)/2} \right\|_{2}
\]
\[
= \left\| \nabla \hat{f}(x_n, N_n^\#) - \nabla f(x_n) \right\|_{2}
\]
From Assumption A 9 and the fact that $C \subset E$ is compact, we know that $\left\| \hat{f}(\cdot, N) - f \right\|_{W_1(C)} \to 0$ as $N \to \infty$.
Since $N_n^\# \to \infty$ as $n \to \infty$, we get that $\left\| \hat{f}(\cdot, N_n^\#) - f \right\|_{W_1(C)} \to 0$ as $n \to \infty$. Further, using the fact that $\{x_n\}_{n \in \mathbb{N}} \subset C$ and the definition of the norm on $W_1(C)$, we get that
\[
\lim_{n \to \infty} \left\| \nabla \hat{f}(x_n, N_n^\#) - \nabla f(x_n) \right\|_{2} = 0
\]
Therefore,
\[
\lim_{n \to \infty} \left\| D_n (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \tilde{Z}_n^T W_n b_n \right\|_2 = 0
\]

Finally, we consider the third term on the right side of (66) for \( n > n^* \).
\[
\left\| D_n (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \tilde{Z}_n^T W_n c_n \right\|_2 \leq \left\| D_n \right\|_2 \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| \tilde{Z}_n^T W_n c_n \right\|_2 = \left( \frac{1}{\delta_n^2} \right) \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| \tilde{Z}_n^T W_n c_n \right\|_2
\]

Using Lemma 3.4 and the definition of \( c_n \), we get
\[
\left\| D_n (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \tilde{Z}_n^T W_n c_n \right\|_2 \leq p \left( \frac{1}{\delta_n^2} \right) \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \times \max_{i \in \{1, \ldots, M_n^I\}} \left\| f(x) \right\|_2 \left\| \nabla f(x) \right\|_2
\]

Again using (68), we get that
\[
\left\| D_n (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \tilde{Z}_n^T W_n c_n \right\|_2 \leq \sqrt{2} p \left( \frac{1}{\delta_n^2} \right) \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \times \max_{i \in \{1, \ldots, M_n^I\}} \left\| f(x) \right\|_2 \left\| \nabla f(x) \right\|_2
\]

Using the fact that \( \max_{i \in \{1, \ldots, M_n^I\}} \left\| f(x) \right\|_2 \left\| \nabla f(x) \right\|_2 \leq K^I_n \), we know that for all \( n \in \mathbb{N} \),
\[
\left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \leq K^I_n
\]

From Assumption A 11 and Lemma 3.8, we know that for all \( n \in \mathbb{N} \),
\[
\left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \left\| (\tilde{Z}_n^T W_n \tilde{Z}_n)^{-1} \right\|_2 \leq K^I_n
\]

Using Assumption A 6 we get
\[
\lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^I\}} \left\| x_n + y_n^i - x_n \right\|_2 = \lim_{n \to \infty} \max_{i \in \{1, \ldots, M_n^I\}} \left\| y_n^i \right\|_2 \leq \delta_n = 0
\]

36
It follows that for any $\delta > 0$, there exists an $n > n^*$ with such that for all $n > n^*, \delta_n < \delta$. Thus, $0 < \|y_n\|_2 \leq \delta$ for all $i = 1, \ldots, M_i^t$ and $n > \tilde{n}$. Thus, we get from (70) that for any $\varepsilon > 0$, there exists $\tilde{n} \in \mathbb{N}$ such that

$$\frac{|f(x_n + y_n^t) - f(x_n) - y_n^i T \nabla f(x_n) - \frac{1}{2} \left( y_n^i T \nabla^2 f(x_n) y_n^i \right)|}{\|y_n\|_2^2} < \varepsilon$$

for $i \in \{1, \ldots, M_i^t\}$ and $n > \tilde{n}$. Consequently,

$$\lim_{n \to \infty} \max_{i \in \{1, \ldots, M_i^t\}} \frac{|f(x_n + y_n^i) - f(x_n) - y_n^i T \nabla f(x_n) - \frac{1}{2} \left( y_n^i T \nabla^2 f(x_n) y_n^i \right)|}{\|y_n\|_2^2} = 0$$

Therefore, we see that

$$\lim_{n \to \infty} \sqrt{2} \ p \ \max_{i \in \{1, \ldots, M_i^t\}} \frac{|f(x_n + y_n^i) - f(x_n) - y_n^i T \nabla f(x_n) - \frac{1}{2} \left( y_n^i T \nabla^2 f(x_n) y_n^i \right)|}{\|y_n\|_2^2} \\ \leq \sqrt{2} \ p \ K_f \ \lim_{n \to \infty} \max_{i \in \{1, \ldots, M_i^t\}} \frac{|f(x_n + y_n^i) - f(x_n) - y_n^i T \nabla f(x_n) - \frac{1}{2} \left( y_n^i T \nabla^2 f(x_n) y_n^i \right)|}{\|y_n\|_2^2} = 0$$

Finally, we show that

$$\lim_{n \to \infty} \sqrt{2} \ p \ \max_{i \in \{1, \ldots, M_i^t\}} \frac{|f(x_n + y_n^i) - f(x_n) - y_n^i T \nabla f(x_n) - \frac{1}{2} \left( y_n^i T \nabla^2 f(x_n) y_n^i \right)|}{\|y_n\|_2^2} = 0$$

As we noted earlier, we get from Assumption A 9, $f$ is twice continuously differentiable on $C$. Further, since $C$ is compact, we get from Lemma 2.3 in ? that there exists $K_{2f} < \infty$ such that

$$\sup_{x, x+y \in C, y \neq 0} \frac{|f(x+y) - f(x) - y \nabla f(x) - \frac{1}{2} y \nabla^2 f(x)y|}{\|y\|_2^2} < K_{2f}$$

Since $\{x_n + y_n^i : i = 1, \ldots, M_n\} \subset C$ for each $n \in \mathbb{N}$ and $\{x_n\}_{n \in \mathbb{N}} \subset C$, it follows that

$$\max_{i \in \{M_i^t+1, \ldots, M_n\}} \frac{|f(x_n + y_n^i) - f(x_n) - y_n^i T \nabla f(x_n) - \frac{1}{2} \left( y_n^i T \nabla^2 f(x_n) y_n^i \right)|}{\|y_n\|_2^2} \leq K_{2f}$$

Also, from Assumption A 11 and Lemma 3.8, we get that

$$\lim_{n \to \infty} \frac{\|\tilde{Z}_n^T W_n \tilde{Z}_n\|_2}{\|\tilde{Z}_n^O T W_n^O \tilde{Z}_n\|_2} = 0$$

Therefore, we finally get

$$\lim_{n \to \infty} \sqrt{2} \ p \ \max_{i \in \{M_i^t+1, \ldots, M_n\}} \frac{|f(x_n + y_n^i) - f(x_n) - y_n^i T \nabla f(x_n) - \frac{1}{2} \left( y_n^i T \nabla^2 f(x_n) y_n^i \right)|}{\|y_n\|_2^2} \leq \sqrt{2} \ p \ K_{2f} \ \lim_{n \to \infty} \frac{\|\tilde{Z}_n^T W_n \tilde{Z}_n\|_2}{\|\tilde{Z}_n^O T W_n^O \tilde{Z}_n\|_2} = 0$$
Thus, we have shown that \( \left\| D_n(\tilde{Z}_n^TW_n\tilde{Z}_n)^{-1}\tilde{Z}_n^TW_nc_n \right\|_2 \to 0 \) as \( n \to \infty \). Hence it holds that

\[
\lim_{n \to \infty} \left\| \left( \begin{array}{c} \nabla_n f(x_n) \\ \nabla^2_n f(x_n) \end{array} \right) - \left( \begin{array}{c} \nabla f(x_n) \\ [\nabla^2 f(x_n)]_{ij} \end{array} \right) \right\|_2 = 0
\]

This in turn gives us

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

In particular, if \( x_n \to \hat{x} \in X \) as \( n \to \infty \), then since \( f \in C_2(C) \),

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

\[
\lim_{n \to \infty} \left\| \nabla^2_n f(x_n) - \nabla^2 f(\hat{x}) \right\|_2 \leq \lim_{n \to \infty} \left\| \nabla^2_n f(x_n) - \nabla^2 f(x_n) \right\|_2 + \lim_{n \to \infty} \left\| \nabla^2 f(x_n) - \nabla^2 f(\hat{x}) \right\|_2 = 0
\]

Thus, we have three results that relate the accuracy of \( \hat{f}(x_n, N_n^0) \), \( \hat{f}_n f(x_n) \) and \( \hat{f}_n^2 f(x_n) \) as respective approximations of \( f(x_n) \), \( \nabla f(x_n) \) and \( \nabla^2 f(x_n) \), to the design points \( \{x_n + y_n^i : i = 1, \ldots, M_n\} \) and the corresponding sample sizes \( N_n^0 \) and \( \{N_n^i : i = 1, \ldots, M_n\} \) and weights \( \{w_n^i : i = 1, \ldots, M_n\} \) used in the determination of \( \hat{f}(x_n, N_n^0) \), \( \hat{f}_n f(x_n) \) and \( \hat{f}_n^2 f(x_n) \) (ultimately used to define the model function \( m_n \)). In the following sections, we discuss various practical schemes that can be used to satisfy the various assumptions made in Lemma 3.1 and Theorems 3.2 and 3.7, which can then be used in our trust region algorithms in Section ?? to construct a sufficiently accurate regression model function for each iteration \( n \).

### 3.1 Smoothness of \( f \) and Convergence of \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \)

Let us first consider Assumptions A 2, A 4 and A 9 made in Lemma 3.1 and Theorems 3.2 and 3.7 respectively. These assumptions place (successively stronger) restrictions the smoothness of the objective function \( f \) and the strength of convergence of the sequence \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \) of sample average functions to the objective function \( f \) as \( N \to \infty \), on a set \( C \subset E \) which contains all the candidate solutions and design points. Next, we obtain sufficient conditions on the smoothness of \( F \) on \( E \) such that Assumptions A 2 and A 4 may be satisfied. Accordingly, we begin by noting a few facts regarding the Clarke generalized gradient defined for Lipschitz continuous functions on \( E \).

Consider a function \( g : E \to \mathbb{R} \) 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.

---

38
**P 3.1.** From Rademacher’s Theorem, there exists a set $\mathcal{U} \subset \mathcal{E}$ with $\mathbf{L}(\mathcal{U}) = 0$ (where $\mathbf{L}$ denotes the Lebesgue measure on $\mathbb{R}^l$) such that $g$ is Frechet differentiable at all $x \in \mathcal{E} \setminus \mathcal{U}$. That is, for any $x \in \mathcal{E} \setminus \mathcal{U}$, 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
\[
\lim_{y \to 0} \frac{g(x + y) - g(x) - y^T \nabla g(x)}{\|y\|_2} = 0
\]

**P 3.2.** Further, it is known that
\[
\sup_{x \in \mathcal{E} \setminus \mathcal{U}} \|\nabla g(x)\|_2 = \sup_{x,y \in \mathcal{E}, x \neq y} \frac{|g(x) - g(y)|}{\|x - y\|_2} = K_g \quad \text{for all } x \in \mathcal{E} \setminus \mathcal{U} \tag{72}
\]
Consequently, if $g \in \mathcal{W}_0(\mathcal{E})$, then
\[
\|g\|_{\mathcal{W}_0(\mathcal{E})} := \sup_{x \in \mathcal{E}} |g(x)| + \sup_{x,y \in \mathcal{E}, x \neq y} \frac{|g(x) - g(y)|}{\|x - y\|_2} = \sup_{x \in \mathcal{E}} |g(x)| + \sup_{x \in \mathcal{E} \setminus \mathcal{U}} \|\nabla g(x)\|_2
\]

**P 3.3.** For any sequence $\{x_n\}_{n \in \mathbb{N}} \subset \mathcal{E} \setminus \mathcal{U}$ such that $x_n \to x \in \mathcal{E} \setminus \mathcal{U}$, we have $\nabla g(x_n) \to \nabla g(x)$ as $n \to \infty$.

Now, for any $x \in \mathcal{E}$, the Clarke generalized gradient $\partial g(x)$ is a set defined as follows
\[
\partial g(x) := \text{conv} \left\{ v \in \mathbb{R}^l : v = \lim_{k \to \infty} \nabla g(x_k) \text{ where } \{x_k\}_{k \in \mathbb{N}} \subset \mathcal{E} \setminus \mathcal{U} \right\} \tag{73}
\]

**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 $g_1, \ldots, g_N \in \mathcal{W}_0(\mathcal{E})$ 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 12.** $F(x, \cdot) : \Xi \to \mathbb{R}$ is $\mathcal{G}$-measurable and $\mathcal{Q}$-integrable for each $x \in \mathcal{E}$.

**A 13.** There exists a $\mathcal{G}$-measurable and $\mathcal{Q}$-integrable function $K : \Xi \to \mathbb{R}_+$, such that for $\mathcal{Q}$-almost all $\zeta$, give any $x, y \in \mathcal{E}$ we have
\[
|F(x, \zeta) - F(y, \zeta)| \leq K(\zeta) \|x - y\|_2 \tag{74}
\]
In particular, there exists a $\mathcal{Q}$-null set $\mathcal{N}_2 \subset \Xi$ such that for $\zeta \notin \mathcal{N}_2$, (74) holds for $K(\zeta) < \infty$.

**A 14.** For any fixed $x \in \mathcal{E}$, there exists a $\mathcal{Q}$-null set $\mathcal{N}_2(x) \subset \Xi$ such that for $\zeta \notin \mathcal{N}_2(x)$, $F(\cdot, \zeta)$ 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 $\left\{f(\cdot, N)\right\}_{N \in \mathbb{N}}$ (for $\mathbf{P}$-almost all $\tilde{\omega} \in \Omega$) and $f$ satisfy the requirements of Assumptions A 2 and A 4.
Lemma 3.9. Suppose $F : \mathcal{E} \times \Xi \rightarrow \mathbb{R}$ satisfies Assumptions A 12 through A 14. Then, given any compact set $D \subset \mathcal{E}$, the following assertions hold.

1. $f$ (as defined in (P)) is continuously differentiable on $\mathcal{E}$, i.e., $\nabla f(x)$ exists and is continuous for each $x \in \mathcal{E}$. Further, for each compact set $D \subset \mathcal{E}$, $f \in \mathcal{C}_1(D)$.

2. For $\mathbb{P}$-almost all $\bar{\omega} \in \Omega$, the sequence $\{\hat{f}(:,N)\}_{N \in \mathbb{N}}$ (where $\hat{f}$ is as defined in (8)), lies in $W_0(D)$.

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

$$\lim_{N \rightarrow \infty} \sup_{x \in D} \left| \hat{f}(x,N) - f(x) \right| = 0 \quad (75)$$

4. For $\mathbb{P}$-almost all $\bar{\omega} \in \Omega$, we have

$$\lim_{N \rightarrow \infty} \sup_{x \in D} \sup_{d \in \partial \hat{f}(x,N)} \|d - \nabla f(x)\|_2 = 0 \quad (76)$$

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

1. Under Assumption A 12 through A 14, 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 $D \subset \mathcal{E}$ is compact, it is easily seen that $f$ and $\nabla f$ are bounded on $D$. Consequently, $\|f\|_{\mathcal{C}_1(D)} < \infty$ and hence we get that $f \in \mathcal{C}_1(D)$.

2. For any $N \in \mathbb{N}$ and $x, y \in D$, we get from (8)

$$\left| \hat{f}(x,N) - \hat{f}(y,N) \right| = \left| \sum_{j=1}^{N} (F(x,\tilde{\zeta}^j) - F(y,\tilde{\zeta}^j)) \right| \leq \sum_{j=1}^{N} \left| F(x,\tilde{\zeta}^j) - F(y,\tilde{\zeta}^j) \right| \quad (77)$$

where $\tilde{\zeta}^j = \zeta^j(\bar{\omega})$ for each $j \in \mathbb{N}$, and $\bar{\omega} \in \Omega$. Consider the set

$$\mathcal{N}_{\Omega}^1 := \bigcup_{j \in \mathbb{N}} (\zeta^j)^{-1}(\mathcal{N}^2_{\Xi}) \quad (78)$$

where $\mathcal{N}^2_{\Xi}$ is defined in Assumption A 13. It is clear that $\mathcal{N}_{\Omega}^1$ is $\mathbb{P}$-null and we get $\{\tilde{\zeta}^j\}_{j \in \mathbb{N}} \subset \Xi \setminus \mathcal{N}_{\Xi}^2$ for all $\bar{\omega} \in \Omega \setminus \mathcal{N}_{\Omega}^1$. Therefore, for $\bar{\omega} \in \Omega \setminus \mathcal{N}_{\Omega}^1$, using (74) in (77) we get that

$$\left| \hat{f}(x,N) - \hat{f}(y,N) \right| \leq \left( \frac{\sum_{j=1}^{N} K(\tilde{\zeta}^j)}{N} \right) \|x - y\|_2 \quad \text{for any } x, y \in D \quad (79)$$

where

$$\left( \frac{\sum_{j=1}^{N} K(\tilde{\zeta}^j)}{N} \right) < \infty$$
Consequently, \( \hat{f}(\cdot, N) \) is continuous on \( \mathcal{D} \) for \( \omega \in \Omega \setminus \mathcal{N}_1 \). Since \( \mathcal{D} \) is compact, we get that \( \sup_{x \in \mathcal{D}} |\hat{f}(x, N)| < K_N \) for some \( K_N < \infty \). Therefore, we finally get that for \( \omega \in \Omega \setminus \mathcal{N}_1 \)

\[
\left\| \hat{f}(\cdot, N) \right\|_{W_0(\mathcal{E})} = \sup_{x \in \mathcal{D}} |\hat{f}(x, N)| + \sup_{x, x + y \in \mathcal{D}, y \neq 0} \left| \frac{\hat{f}(x + y, N) - \hat{f}(x, N)}{\|y\|_2} \right| \leq K_N + \left( \frac{\sum_{j=1}^{N} K(\zeta_j)}{N} \right) < \infty
\]

Hence \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \subseteq W_0(\mathcal{D}) \) for \( \mathbf{P} \)-almost all \( \omega \in \Omega \).

3. It is clear from (74) in Assumption A 13 that for \( \mathbf{Q} \)-almost all \( \zeta \), the function \( F(\cdot, \zeta) \) is continuous on \( \mathcal{D} \). Now, consider some \( x^* \in \mathcal{D} \) such that \( \mathbb{E}_{\mathbf{Q}}[F(x^*, \zeta)] < \infty \). Such an \( x^* \) exists from Assumption A 12. Since \( \mathcal{D} \) is compact, there exists \( K_D < \infty \) such that \( \sup_{x \in \mathcal{D}} \|x - x^*\|_2 < K_D \). Therefore, we get using (74) that for \( \mathbf{Q} \)-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 12 we get that \( |F(x^*, \zeta)| \) is \( \mathbf{Q} \)-integrable and from Assumption A 13, we know that \( K(\zeta) \) is \( \mathbf{Q} \)-integrable. Therefore the family \( \{ |F(x, \zeta)| : x \in \mathcal{D} \} \) is dominated by a \( \mathbf{Q} \)-integrable function. Thus using, Lemma A1 (on page 67) in ?, we get that (75) holds.

4. Finally, we show that (76) holds. Note that the statement of (76) is well-defined since we have already shown that \( f \in C_1(\mathcal{D}) \subseteq C_1(\mathcal{E}) \) and that there exists a \( \mathbf{P} \)-null set \( \mathcal{N}_1^\mathcal{D} \subseteq \Omega \) defined as in (78), such that for all \( \omega \notin \mathcal{N}_1^\mathcal{D} \), \( \{\hat{f}(\cdot, N)\}_{N \in \mathbb{N}} \subseteq W_0(\mathcal{D}) \).

In order to show (76), we define for any \( \bar{x} \in \mathcal{D} \) and \( \bar{\delta} > 0 \),

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

(80)

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

\[
G_k^\bar{x}(\zeta) := \sup_{x \in \mathcal{B}(\bar{x}, \delta_k)} \sup_{d \in \partial F(x, \zeta)} \|d - \nabla F(\bar{x}, \zeta)\|_2
\]

(81)

For each \( k \in \mathbb{N} \), we know that \( \mathcal{B}(\bar{x}, \delta_k) \subseteq C \subseteq \mathcal{E} \). Also, we know from Assumption A 13 that for \( \zeta \notin \mathcal{N}_2^\mathcal{E} \) (where \( \mathbf{Q}(\mathcal{N}_2^\mathcal{E}) = 0 \)), (74) holds for \( x, y \in \mathcal{E} \) and consequently we have

\[
\sup_{x, y \in \mathcal{D}, x \neq y} \left| \frac{F(x, \zeta) - F(y, \zeta)}{\|x - y\|_2} \right| \leq K(\zeta) \leq \infty
\]

Therefore, for each \( x \in \mathcal{D} \), we can define a a Clarke generalized gradient \( \partial F(x, \zeta) \) as in (73). Further, from Assumption A 14, we get that there exists a \( \mathbf{Q} \)-null set \( \mathcal{N}_2^\mathcal{D}(\bar{x}) \) such that for \( \zeta \notin \mathcal{N}_2^\mathcal{D}(\bar{x}) \), \( F(\cdot, \zeta) \) is
As a consequence, we get the following two properties of the sequence\( G_k^2 \).

Thus, we have shown the converse inequality and hence (82) holds for each \( k \leq N \setminus \left( N_k^2 \cup N_{\tilde{x}}^2 (\tilde{x}) \right) \).

From Assumption A 13 and Property P 3.1, we know that if \( \zeta \notin N_1^2 \), \( \partial F(x, \zeta) = \{ \nabla F(x, \zeta) \} \) for \( x \in D \setminus U_D(\zeta) \) where \( L(U_D(\zeta)) = 0 \). Using this we show next that for each \( \zeta \in \Xi \setminus \left( N_k^2 \cup N_{\tilde{x}}^2 (\tilde{x}) \right) \) and \( k \leq N \),

\[
G_k^2(\zeta) = \sup_{x \in B(\tilde{x}, \delta_k) \setminus U_D(\zeta)} \| \nabla F(x, \zeta) - \nabla F(\tilde{x}, \zeta) \|_2 \tag{82}
\]

From (81) and Property P 3.4, it is easily seen that

\[
G_k^2(\zeta) \geq \sup_{x \in B(\tilde{x}, \delta_k) \setminus U_D(\zeta)} \sup_{d \in \partial F(x, \zeta)} \| d - \nabla F(\tilde{x}, \zeta) \|_2 = \sup_{x \in B(\tilde{x}, \delta_k) \setminus U_D(\zeta)} \| \nabla F(x, \zeta) - \nabla F(\tilde{x}, \zeta) \|_2
\]

Therefore, we only have to show the converse inequality to prove (82). To do this, first we note that for any set \( C \subseteq \mathbb{R}^l \) and \( d^* \subseteq \mathbb{R}^l \), we have

\[
\sup_{d \in C} \| d - d^* \|_2 = \sup_{x \in \text{conv}(C)} \| d - d^* \|_2 \tag{83}
\]

Let us fix \( \zeta \in \Xi \setminus \left( N_k^2 \cup N_{\tilde{x}}^2 (\tilde{x}) \right) \) and \( k \leq N \). Then, for each \( x \in B(\tilde{x}, \delta_k) \), using (83) and the definition of the Clarke generalized gradient in (73), we get

\[
\sup_{d \in \partial F(x, \zeta)} \| d - \nabla F(\tilde{x}, \zeta) \|_2 = \sup \left\{ \| d - \nabla F(\tilde{x}, \zeta) \|_2 : d = \lim_{j \to x} \nabla F(x_j, \zeta), \{ x_j \}_{j \leq N} \subseteq B(\tilde{x}, \delta_k) \setminus U_D(\zeta) \right\} \tag{84}
\]

Now, consider some \( d \in \mathbb{R}^l \) such that \( d = \lim_{x_j \to x} \nabla F(x_j, \zeta) \) for some sequence \( \{ x_j \}_{j \leq N} \subseteq B(\tilde{x}, \delta_k) \setminus U_D(\zeta) \). We get for each \( j \leq N \),

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

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

\[
\| d - \nabla F(\tilde{x}, \zeta) \|_2 \leq \sup_{x \in B(\tilde{x}, \delta_k) \setminus U_D(\zeta)} \| \nabla F(x, \zeta) - \nabla F(\tilde{x}, \zeta) \|_2
\]

Consequently, using (84),

\[
\sup_{d \in \partial F(x, \zeta)} \| d - \nabla F(\tilde{x}, \zeta) \|_2 \leq \sup_{x \in B(\tilde{x}, \delta_k) \setminus U_D(\zeta)} \| \nabla F(x, \zeta) - \nabla F(\tilde{x}, \zeta) \|_2
\]

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

\[
G_k^2(\zeta) := \sup_{x \in B(\tilde{x}, \delta_k)} \sup_{d \in \partial F(x, \zeta)} \| d - \nabla F(\tilde{x}, \zeta) \|_2 \leq \sup_{x \in B(\tilde{x}, \delta_k) \setminus U_D(\zeta)} \| \nabla F(x, \zeta) - \nabla F(\tilde{x}, \zeta) \|_2
\]

Thus, we have shown the converse inequality and hence (82) holds for each \( k \leq N \) and \( \zeta \in \Xi \setminus \left( N_k^2 \cup N_{\tilde{x}}^2 (\tilde{x}) \right) \).

As a consequence, we get the following two properties of the sequence \( \{ G_k^2 \}_{k \leq N} \).
Therefore, using the Lebesgue Dominated Convergence Theorem, we get that
\[ G^x_k(\zeta) \leq 2 \sup_{x \in B(\hat{x}, \hat{\delta}_k)} \| \nabla F(x, \zeta) \|_2 \leq 2 \sup_{x \in D(\hat{\delta}_k)} \| \nabla F(x, \zeta) \|_2 \leq 2K(\zeta) \]

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

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

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

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

Therefore, using the Lebesgue Dominated Convergence Theorem, we get that
\[ \lim_{k \to \infty} \mathbb{E}_Q \left[ G^x_k(\zeta) \right] = 0 \]

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

Now, we define the \( P \)-null set \( \Lambda^2_\Omega(\tilde{x}) \subset \Omega \) as
\[ \Lambda^2_\Omega(\tilde{x}) := \bigcup_{j \in \mathbb{N}} (\zeta^j)^{-1}(\Lambda^2_\Omega(\tilde{x})) \quad (85) \]

If \( \tilde{\omega} \notin \Lambda^2_\Omega(\tilde{x}) \), then by definition \( \tilde{\zeta}^j := \zeta^j(\tilde{\omega}) \notin \Lambda^2_\Omega(\tilde{x}) \) for each \( j \in \mathbb{N} \). Recall that we also analogously defined the \( P \)-null set \( \Lambda^1_\Omega \subset \Omega \) in (78). Next, we consider the sample average functions \( \{ \hat{f}(\cdot, N) \}_{N \in \mathbb{N}} \) for \( \tilde{\omega} \in \Omega \setminus (\Lambda^1_\Omega \cup \Lambda^2_\Omega(\tilde{x})) \). First of all it is easily seen that for \( \tilde{\omega} \in \Omega \setminus (\Lambda^1_\Omega \cup \Lambda^2_\Omega(\tilde{x})) \), \( \hat{f}(\cdot, N) \) is differentiable at \( \hat{x} \) for each \( N \in \mathbb{N} \) and
\[ \nabla \hat{f}(\hat{x}, N) = \frac{1}{N} \sum_{j=1}^{N} \nabla F(\hat{x}, \tilde{\zeta}^j) \]

Further, for any \( x \in D, N \in \mathbb{N} \) and \( \tilde{\omega} \in \Omega \setminus (\Lambda^1_\Omega \cup \Lambda^2_\Omega(\tilde{x})) \), it is easily seen from Property P 3.5.
\[ \partial \hat{f}(x, N) = \partial \left\{ \frac{1}{N} \sum_{j=1}^{N} F(\cdot, \tilde{\zeta}^j) \right\}(x) \subseteq \frac{1}{N} \sum_{j=1}^{N} \partial F(x, \tilde{\zeta}^j) \]

Using these two observations, we get for \( \tilde{\omega} \in \Omega \setminus (\Lambda^1_\Omega \cup \Lambda^2_\Omega(\tilde{x})) \), \( k \in \mathbb{N} \) and \( x \in B(\hat{x}, \hat{\delta}_k) \),
\[ \sup_{d \in \partial \hat{f}(x, N)} \left\| d - \nabla \hat{f}(\hat{x}, N) \right\|_2 \leq \sup_{d \in \partial \hat{f}(x, N)} \left\| d - \nabla \hat{f}(\hat{x}, N) \right\|_2 = \sup \left\{ \left\| d - \nabla \hat{f}(\hat{x}, N) \right\|_2 : d \in \partial \left\{ \frac{1}{N} \sum_{j=1}^{N} F(\cdot, \tilde{\zeta}^j) \right\}(x) \right\} \]
\[ \leq \sup \left\{ \left\| d - \nabla \hat{f}(\hat{x}, N) \right\|_2 : d \in \frac{1}{N} \sum_{j=1}^{N} \partial F(x, \tilde{\zeta}^j) \right\} \]
\[ = \sup \left\{ \frac{1}{N} \sum_{j=1}^{N} \left\| d^j - \nabla F(\hat{x}, \tilde{\zeta}^j) \right\|_2 : d^j \in \partial F(x, \tilde{\zeta}^j) \text{ for } j = 1, \ldots, N \right\} \]
\[ \leq \sup \left\{ \frac{1}{N} \sum_{j=1}^{N} \left\| d^j - \nabla F(\hat{x}, \tilde{\zeta}^j) \right\|_2 : d^j \in \partial F(x, \tilde{\zeta}^j) \text{ for } j = 1, \ldots, N \right\} \]
\[ = \frac{1}{N} \sum_{j=1}^{N} \sup_{d \in \partial F(x, \tilde{\zeta}^j)} \left\| d - \nabla F(\hat{x}, \tilde{\zeta}^j) \right\|_2 \]
Therefore, we get that for $\bar{\omega} \in \Omega \setminus \left( \mathcal{N}_m^1 \bigcup \mathcal{N}_2^2(\hat{x}) \right)$ and $k \in \mathbb{N}$

$$
\sup_{x \in B(\hat{x}, \delta_k)} \sup_{d \in \partial f(x,N)} \left\| d - \nabla \hat{f}(\hat{x}, N) \right\|_2 = \sup_{x \in B(\hat{x}, \delta_k)} \left\{ \frac{1}{N} \sum_{j=1}^{N} \sup_{d \in \partial f(x, \tilde{\zeta}^j)} \left\| d - \nabla F(\hat{x}, \tilde{\zeta}^j) \right\|_2 \right\}
\leq \frac{1}{N} \sum_{j=1}^{N} \left\{ \sup_{x \in B(\hat{x}, \delta_k)} \sup_{d \in \partial f(x, \tilde{\zeta}^j)} \left\| d - \nabla F(\hat{x}, \tilde{\zeta}^j) \right\|_2 \right\} = \frac{1}{N} \sum_{j=1}^{N} G_k^z(\tilde{\zeta}^j)
$$

Now, consider the right side of the last inequality given above. Since $\{\tilde{\zeta}^j\}_{j \in \mathbb{N}}$ is an i.i.d. sequence, we get from the strong law large numbers that $\frac{1}{N} \sum_{j=1}^{N} G_k^z(\tilde{\zeta}^j)$ converges to $E_Q \left[ G_k^z(\zeta) \right] = \mathbb{E}_P \left[ G_k^z(\zeta(\bar{\omega})) \right]$ for $P$-almost all $\bar{\omega}$. That is, there exists a $P$-null set $\mathcal{N}_m^1(\hat{x}, k) \subset \Omega$ such that for $\bar{\omega} \in \Omega \setminus \left( \mathcal{N}_m^1 \bigcup \mathcal{N}_2^2(\hat{x}) \bigcup \mathcal{N}_2^2(\hat{x}, k) \right)$, the following statement holds. For each $\epsilon > 0$, there exists $N(\hat{x}, k, \bar{\omega}, \epsilon)$ such that

$$
\frac{\sum_{j=1}^{N} G_k^z(\tilde{\zeta}^j)}{N} - E_Q \left[ G_k^z(\zeta) \right] < \epsilon \quad \text{for all } N \geq N(\hat{x}, k, \bar{\omega}, \epsilon)
$$

Therefore, we finally get that given any $\epsilon > 0$ and $\hat{x} \in \mathcal{D}$, for each $k > k(\hat{x}, \epsilon)$, there exists a $P$-null set $\mathcal{N}_m^1(\hat{x}, k)$ such that if $\bar{\omega} \in \Omega \setminus \left( \mathcal{N}_m^1 \bigcup \mathcal{N}_2^2(\hat{x}) \bigcup \mathcal{N}_2^2(\hat{x}, k) \right)$, then

$$
\sup_{x \in B(\hat{x}, \delta_k)} \sup_{d \in \partial f(x,N)} \left\| d - \nabla \hat{f}(\hat{x}, N) \right\|_2 < 2\epsilon \quad \text{for all } N > N(\hat{x}, k, \bar{\omega}, \epsilon) \quad (86)
$$

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

$$
\sup \{ \| \nabla f(x) - \nabla f(y) \|_2 : x, y \in \mathcal{D}, \| y - x \|_2 \leq \delta(\epsilon) \} < \epsilon \quad (87)
$$

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

$$
\mathcal{D} \subseteq \bigcup_{j=1}^{m^i} B(\hat{x}_j^i, k_x^i) \quad (88)
$$

For each $j = 1, \ldots, m^i$, the following statements clearly hold.

- From Assumption A 13, if $\bar{\omega} \in \Omega \setminus \mathcal{N}_m^2(\hat{x}_j^i)$, then $\nabla \hat{f}(\hat{x}_j^i, N)$ exists for each $N \in \mathbb{N}$.

- From the Strong Law of Large Numbers, there exists a $P$-null set $\mathcal{N}_m^2(\hat{x}_j^i)$ such that if $\bar{\omega} \in \Omega \setminus \left( \mathcal{N}_m^2(\hat{x}) \bigcup \mathcal{N}_2^2(\hat{x}_j^i) \right)$, then for any $\epsilon > 0$, there exists $\bar{N}(\hat{x}_j^i, \bar{\omega}, \epsilon)$ such that

$$
\left\| \nabla \hat{f}(\hat{x}_j^i, N) - \nabla f(\hat{x}_j^i) \right\|_2 < \epsilon \quad \text{for all } N > \bar{N}(\hat{x}_j^i, \bar{\omega}, \epsilon) \quad (89)
$$
Let us define
\[ N(i) := \bigcup_{j=1}^{m} \left\{ N_{\Omega}^0 \cup N_{\Omega}^2(\bar{x}_j^i) \cup N_{\Omega}^3(\bar{x}_j^i, k_{\bar{x}_j^i}) \cup N_{\Omega}^4(\bar{x}_j^i) \right\} \]
\[ N(i, \bar{\omega}) := \max_{j=1, \ldots, m} \max \left\{ N(\bar{x}_j^i, k_{\bar{x}_j^i}, \bar{\omega}, (1/i)) \right\} \]

It is easily seen that \( P(N_\Omega(i)) = 0 \).

Now, for any \( \bar{\omega} \in \Omega \setminus N_\Omega(i) \) and \( x \in D \), we know from (88) that there exists \( j \in \{1, \ldots, m\} \) such that \( x \in \mathcal{B}(\bar{x}_j^i, \delta_{\bar{x}_j^i}) \). Therefore, for each \( N > N(i, \bar{\omega}) \left( \geq N(\bar{x}_j^i, k_{\bar{x}_j^i}, \bar{\omega}, (1/i)) \right) \), we get from (86) that
\[
\sup_{d \in \partial f(x, N)} \left\| d - \nabla \hat{f}(\bar{x}_j^i, N) \right\|_2 < \frac{2}{i}.
\]

Similarly, for each \( N > N(i, \bar{\omega}) \left( \geq N(\bar{x}_j^i, \bar{\omega}, (1/i)) \right) \), we have from (89) that
\[
\left\| \nabla \hat{f}(\bar{x}_j^i, N) - \nabla f(\bar{x}_j^i) \right\|_2 < (1/i).
\]

Thus, since we chose \( k_{\bar{x}_j^i} \), such that \( \tilde{\delta}_{k_{\bar{x}_j^i}} < \tilde{\delta}(1/i) \), we get from (87) that
\[
\left\| f(x) - \nabla f(\bar{x}_j^i) \right\|_2 < (1/i).
\]

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

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

Finally, if we set \( N_\Omega := \bigcup \{ N_\Omega(i) : i \in \mathbb{N} \} \), it is clear that \( P(N_\Omega) = 0 \). Then, for \( \bar{\omega} \in \Omega \setminus N_\Omega \) we get that for each \( i \in \mathbb{N} \), there exists \( N(i, \bar{\omega}) \) such that for \( N > N(i, \bar{\omega}) \), (90) holds. Therefore we get that for \( P \)-almost all \( \bar{\omega} \),
\[
\lim_{N \to \infty} \sup_{x \in D} \sup_{d \in \partial f(x, N)} \left\| d - \nabla f(x) \right\|_2 = 0.
\]

From Lemma 3.9, we get the following corollary.

**Corollary 3.10.** Let Assumptions A 12 through A 14 hold. Then, we get
\[
\lim_{N \to \infty} \left\| \hat{f}(\cdot, N) - f \right\|_{W_0(D)} = 0.
\]

Consequently, there exists \( K_f < \infty \), such that \( \| f \|_{W_1(D)} < K_f \) and \( \| \hat{f}(\cdot, N) \|_{W_1(D)} < K_f \) for each \( N \in \mathbb{N} \).

**Proof.** Since all the assumptions of Lemma 3.9 hold, we see \( \{ \hat{f}(\cdot, N) \}_{N \in \mathbb{N}} \in W_0(D) \) for \( P \)-almost all \( \bar{\omega} \). Also, \( f \in C_1(D) \subset W_0(D) \) Therefore, the statement of Corollary 3.10 is well defined.
Now, from Lemma 3.9, there exists a set $N_{1\Omega} \subset \Omega$, such that for all $\bar{\omega} \in \Omega \setminus N_{1\Omega}$, \( \{ \bar{f}(\cdot, N) \}_{n \in \mathbb{N}} \subset \mathbb{W}_0(\mathcal{D}) \) and (75), (76) hold. Consider any such $\bar{\omega} \in \Omega \setminus N_{1\Omega}$. From Rademacher’s theorem, we know that for each $N \in \mathbb{N}$, there exists a set $U_N(\bar{\omega}) \subset \mathcal{D}$ with $\mathbf{L}(U_N(\bar{\omega})) = 0$, such that $\nabla \bar{f}(\cdot, N)$ exists for all $x \in \mathcal{D} \setminus U_N(\bar{\omega})$. Using Property P 3.4, we get that

\[
\left\| \bar{f}(\cdot, N) - f \right\|_{\mathbb{W}_0(\mathcal{D})} = \sup_{x \in \mathcal{D}} |\bar{f}(x, N) - f(x)| + \sup_{x \in \mathcal{D} \setminus U_N(\bar{\omega})} \left\| \nabla \bar{f}(x, N) - \nabla f(x) \right\|_2 \tag{91}
\]

It follows from (75) that

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

Also, for each $N \in \mathbb{N}$,

\[
\sup_{x \in \mathcal{D}} \sup_{d \in \partial \bar{f}(x,N)} \left\| d - \nabla \bar{f}(x) \right\|_2 \geq \sup_{x \in \mathcal{D} \setminus U_N(\bar{\omega})} \sup_{d \in \partial \bar{f}(x,N)} \left\| d - \nabla f(x) \right\|_2
\]

But we know that for $x \in \mathcal{D} \setminus U_N(\bar{\omega})$, $\bar{f}(\cdot, N)$ is differentiable and hence, $\partial \bar{f}(\cdot, N) = \{ \nabla \bar{f}(\cdot, N) \}$. Thus,

\[
\sup_{x \in \mathcal{D} \setminus U_N(\bar{\omega})} \sup_{d \in \partial \bar{f}(x,N)} \left\| d - \nabla \bar{f}(x) \right\|_2 = \sup_{x \in \mathcal{D} \setminus U_N(\bar{\omega})} \left\| \nabla \bar{f}(x, N) - \nabla f(x) \right\|_2
\]

Therefore, we get from (76) that

\[
0 = \lim_{N \to \infty} \sup_{x \in \mathcal{D}} \sup_{d \in \partial \bar{f}(x,N)} \left\| d - \nabla f(x) \right\|_2 \geq \lim_{N \to \infty} \sup_{x \in \mathcal{D} \setminus U_N(\bar{\omega})} \left\| \nabla \bar{f}(x, N) - \nabla f(x) \right\|_2
\]

Thus, we have for all $\omega \in \Omega \setminus N_{1\Omega}$,

\[
\lim_{N \to \infty} \left\| \bar{f}(\cdot, N) - f \right\|_{\mathbb{W}_0(\mathcal{D})} = \lim_{N \to \infty} \sup_{x \in \mathcal{D}} |\bar{f}(x, N) - f(x)| + \lim_{N \to \infty} \sup_{x \in \mathcal{D} \setminus U_N(\bar{\omega})} \left\| \nabla \bar{f}(x, N) - \nabla f(x) \right\|_2 = 0
\]

Therefore, $\left\| \bar{f}(\cdot, N) - f \right\|_{\mathbb{W}_0(\mathcal{D})} \to 0$ as $N \to \infty$ for $\mathbf{P}$-almost all $\omega$. Hence $\left\| \bar{f}(\cdot, N) \right\|_{\mathbb{W}_0(\mathcal{D})} \to \left\| f \right\|_{\mathbb{W}_0(\mathcal{D})}$ as $N \to \infty$ and consequently, the sequence $\left\{ \left\| \bar{f}(\cdot, N) \right\|_{\mathbb{W}_0(\mathcal{D})} \right\}_{N \in \mathbb{N}}$ is bounded. Therefore, there exists $K_f \in (\left\| f \right\|_{\mathbb{W}_0(\mathcal{D})}, \infty)$ such that $\left\| \bar{f}(\cdot, N) \right\|_{\mathbb{W}_0(\mathcal{D})} < K_f$ for each $N \in \mathbb{N}$. \( \square \)

From Lemma 3.9 and Corollary 3.10, it is clear that both Assumptions A 2 and A 4 are satisfied if Assumptions A 12 through 14 hold. For the rest of this paper, we will assume that we have chosen a particular $\bar{\omega} \in \Omega$ such that Assumptions A 2 and A 4 hold.

Note that the results of Lemma 3.9 and Corollary 3.10 do not satisfy Assumption A 9 which requires that $f \in C_2(\mathcal{D})$ and $\left\| \bar{f}(\cdot, N) - f \right\|_{\mathbb{W}_1(\mathcal{D})} \to 0$ as $N \to \infty$ for any compact set $\mathcal{D} \subset \mathcal{E}$. We do not consider the conditions required to satisfy Assumption A 9, since we will not require the Hessian approximation $\nabla^2 f(x_n)$ to get progressively more accurate as $n \to \infty$, in order to show convergence of our trust region algorithms. We will not even require that $f \in C_2(\mathcal{E})$, i.e., that $\nabla^2 f(x)$ exist for any $x \in \mathcal{D}$. Indeed the only assumption
we will make regarding the quadratic component of our model functions, is Assumption A 8 which ensures that \( \{ \hat{\nabla}_n^2 f(x_n) \} \) used remains bounded.

However, we will present the rest of our analysis assuming that we use a quadratic model function as in (9). Further, in the next section, we will discuss methods to pick the design points and the corresponding sample sizes and weights not only to satisfy the assumptions in Theorem 3.2 but also to satisfy the remaining assumptions (A 10 and A 11) in Theorem 3.7. Thus, in the event that \( f \) and \( \{ \hat{f}(\cdot, N) \} \) are known to satisfy the Assumption A 9, such methods may be used to find a good approximation of the Hessian \( \nabla^2 f(x_n) \) and in turn, a more accurate model function \( m_n \) for each \( n \in \mathbb{N} \).

### 3.2 Picking the Design Points, Sample Sizes and Weights

In this section, we discuss methods to pick the number of design points \( M_n \), the design points \( \{ x_n + y^i_n : i = 1, \ldots, M_n \} \), the sample sizes \( \{ N^i_n : i = 1, \ldots, M_n \} \) and the weights \( \{ w^i_n : i = 1, \ldots, M_n \} \) for each \( n \in \mathbb{N} \), so as to satisfy the assumptions made in Theorems 3.2 and 3.7. Here we will assume that for each \( n \in \mathbb{N} \), we have knowledge of \( x_n \in \mathcal{X} \), the design region radius \( \delta_n > 0 \) and the sample size \( N^0_n \). Further, we will assume that \( \{ N^0_n \} \) satisfies

\[
\lim_{n \to \infty} N^0_n = \infty \quad \text{and} \quad N^0_{n+1} \geq N^0_n \quad \text{for each} \quad n \in \mathbb{N} \tag{92}
\]

and that

\[
\lim_{n \to \infty} \delta_n = 0 \quad \text{and} \tag{93}
\]

When we describe our trust region algorithms in Section ??, we will show how \( \delta_n \) and \( N^0_n \) can be adaptively set for each iteration \( n \), such that (92) and (93) hold.

It is clear that for each \( n \in \mathbb{N} \), once the \( M_n \) design points \( \{ x_n + y^i_n : i = 1, \ldots, M_n \} \) and the corresponding sample sizes \( \{ N^i_n : i = 1, \ldots, M_n \} \) are set, in order to construct the regression model function \( m_n \) as in (9), we need to evaluate for each \( n \in \mathbb{N} \), \( \hat{f}(x_n, N^0_n) \) and also \( \hat{f}(x_n, N^i_n) \) and \( \hat{f}(x_n + y^i_n, N^i_n) \) for \( i = 1, \ldots, M_n \). Consequently, we require \( \max\{ N^0_n, N^1_n, \ldots, N^M_n \} + \sum_{i=1}^{M_n} N^i_n \) evaluations of \( F \) for each \( n \in \mathbb{N} \). Now, keeping in mind that such a regression model function will eventually be used in a trust region algorithm to solve (P), recall that in Section 1, we made the assumption (A 1) that the evaluation of the function \( F \) is computationally expensive for any \( x \in \mathbb{R}^l \) and \( \zeta \in \Xi \). Thus, the methods we develop to pick the design points and sample sizes for each \( n \in \mathbb{N} \), must ensure not only that assumptions of Theorem 3.2 or Theorem 3.7 are satisfied, but also that the number of evaluations of \( F \) required to subsequently evaluate \( \tilde{\nabla}_n f(x_n) \) and \( \tilde{\nabla}_n^2 f(x_n) \) is minimized.

We begin in Section 3.2.1 by developing procedures to pick the inner and outer design points for each \( n \in \mathbb{N} \). Subsequently, we consider methods to pick the corresponding sample sizes and weights in Sections
3.2.1 Design points

It is easily seen that the conditions imposed on the design points and weights in Theorems 3.2 and 3.7, occur in (35) through (37) in Assumption A 7 and in (57) through (59) in Assumption A 11 respectively. In this section, we provide methods to pick design points \( \{x_n + y_i^T n : i = 1, \ldots, M_n\} \) for each \( n \in \mathbb{N} \) such that under certain assumptions on the weights \( \{w_i^T n : i = 1, \ldots, M_n\} \), (35) and (36) in Assumption A 7 and (57) and (58) in Assumption A 11 are satisfied. Later, in Section 3.2.3, we deal with methods to assign an appropriate weight to each design points so as to satisfy the assumptions made in this section and the conditions (37) and (59).

First, we consider the conditions (36) and (58) in Assumptions A 7 and A 11 respectively. Let us start by stating some well known and relevant properties of matrices. Consider a matrix \( Y \in \mathbb{R}^{m \times l} \) given by

\[
Y := \begin{pmatrix}
y_1^T \\
y_2^T \\
\vdots \\
y_m^T
\end{pmatrix}
\]

where \( y_i^T \in \mathbb{R}^l \) for each \( i = 1, \ldots, m \) (94)

1. The matrix \( Y^T Y = \sum_{i=1}^m y_i y_i^T \in \mathbb{R}^{l \times l} \) is symmetric and positive semidefinite.

2. We get from the Courant-Fischer eigenvalue characterization that

\[
\|Y^T Y\|_2 = \lambda_{\text{max}}(Y^T Y) = \max_{x \in \mathbb{R}^l, \|x\|_2 = 1} x^T (Y^T Y) x = \max_{x \in \mathbb{R}^l, \|x\|_2 = 1} \sum_{i=1}^m \left( y_i^T x \right)^2
\]

and

\[
\lambda_{\text{min}}(Y^T Y) = \min_{x \in \mathbb{R}^l, \|x\|_2 = 1} x^T (Y^T Y) x = \min_{x \in \mathbb{R}^l, \|x\|_2 = 1} \sum_{i=1}^m \left( y_i^T x \right)^2
\]

If \( Y^T Y \) is positive definite, then \( \lambda_{\text{min}}(Y^T Y) > 0 \), \((Y^T Y)^{-1}\) exists and \( \|(Y^T Y)^{-1}\|_2 = (1/\lambda_{\text{min}}(Y^T Y)) \).

3. The quantity \( \|Y^T Y\|_2 \|(Y^T Y)^{-1}\|_2 = (\lambda_{\text{max}}(Y^T Y)/\lambda_{\text{min}}(Y^T Y)) \) is called the (two-norm) condition number of the matrix \( Y^T Y \). We have

\[
1 \leq \|Y^T Y\|_2 \|(Y^T Y)^{-1}\|_2 \leq \infty
\]

with the convention that \( \|Y^T Y\|_2 \|(Y^T Y)^{-1}\|_2 := \infty \) whenever \( \lambda_{\text{min}}(Y^T Y) = 0 \) (and \( Y^T Y^{-1} \) does not exist).
Now, consider the matrix \((\hat{Y}_n^T W_n^T \hat{Y}_n^I)\) occurring in (36). Recall that \(W_n^I = \text{diag} \left( w_1^n, \ldots, w_{M_n^I}^n \right) \) where \(w_i^n > 0\) is the weight assigned to inner design point \(x_n + y_i^n\) for each \(i = 1, \ldots, M_n^I\) and \(n \in \mathbb{N}\). It is easily seen that for each \(n \in \mathbb{N}\),

\[
(\hat{Y}_n^T W_n^I \hat{Y}_n^I) = \left( \sqrt{W_n^I} \hat{Y}_n^I \right)^T \left( \sqrt{W_n^I} \hat{Y}_n^I \right)
\]

Therefore, using (95) and (96), we get

\[
\left\| (\hat{Y}_n^T W_n^I \hat{Y}_n^I) \right\|_2 = \lambda_{\max} \left( \left( \sqrt{W_n^I} \hat{Y}_n^I \right)^T \left( \sqrt{W_n^I} \hat{Y}_n^I \right) \right)
\]

\[
= \max_{x \in \mathbb{R}^I, \|x\|_2 = 1} \sum_{i=1}^{M_n^I} w_i^n (\hat{y}_i^n)^T x
\]

\[
\leq \left( \max_{i=1, \ldots, M_n^I} w_i^n \right) \left\{ \max_{x \in \mathbb{R}^I, \|x\|_2 = 1} \sum_{i=1}^{M_n^I} (\hat{y}_i^n)^T x \right\}
\]

\[
= \left( \max_{i=1, \ldots, M_n^I} w_i^n \right) \left\| (\hat{Y}_n^T \hat{Y}_n^I) \right\|_2
\]

and

\[
\lambda_{\min} \left( \left( \sqrt{W_n^I} \hat{Y}_n^I \right)^T \left( \sqrt{W_n^I} \hat{Y}_n^I \right) \right) = \min_{x \in \mathbb{R}^I, \|x\|_2 = 1} \sum_{i=1}^{M_n^I} w_i^n (\hat{y}_i^n)^T x
\]

\[
\geq \left( \min_{i=1, \ldots, M_n^I} w_i^n \right) \left\{ \min_{x \in \mathbb{R}^I, \|x\|_2 = 1} \sum_{i=1}^{M_n^I} (\hat{y}_i^n)^T x \right\}
\]

\[
= \left( \min_{i=1, \ldots, M_n^I} w_i^n \right) \lambda_{\min} \left( (\hat{Y}_n^T \hat{Y}_n^I) \right)
\]

which gives us

\[
\left\| (\hat{Y}_n^T W_n^I \hat{Y}_n^I)^{-1} \right\|_2 = \frac{1}{\lambda_{\min} \left( \left( \sqrt{W_n^I} \hat{Y}_n^I \right)^T \left( \sqrt{W_n^I} \hat{Y}_n^I \right) \right)} \leq \frac{1}{\min_{i=1, \ldots, M_n^I} w_i^n} \left\| (\hat{Y}_n^T \hat{Y}_n^I)^{-1} \right\|_2
\]

Therefore, it is clear that

\[
\left\| (\hat{Y}_n^T W_n^I \hat{Y}_n^I) \right\|_2 \left\| (\hat{Y}_n^T W_n^I \hat{Y}_n^I)^{-1} \right\|_2 \leq \left( \frac{\max_{i=1, \ldots, M_n^I} w_i^n}{\min_{i=1, \ldots, M_n^I} w_i^n} \right) \left\| (\hat{Y}_n^T \hat{Y}_n^I) \right\|_2 \left\| (\hat{Y}_n^T \hat{Y}_n^I)^{-1} \right\|_2
\]

(97)

Similarly, it is easy to show that

\[
\left\| (\hat{Z}_n^T W_n^I \hat{Z}_n^I) \right\|_2 \left\| (\hat{Z}_n^T W_n^I \hat{Z}_n^I)^{-1} \right\|_2 \leq \left( \frac{\max_{i=1, \ldots, M_n^I} w_i^n}{\min_{i=1, \ldots, M_n^I} w_i^n} \right) \left\| (\hat{Z}_n^T \hat{Z}_n^I) \right\|_2 \left\| (\hat{Z}_n^T \hat{Z}_n^I)^{-1} \right\|_2
\]

(98)

Now, in order to satisfy (36) and (58) we will obtain separate bounds on each of the terms occurring in the products on the right sides of (97) and (98) respectively.

First, we will make the following assumption in this section regarding the weights \(\{w_i^n : i = 1, \ldots, M_n^I\}\) for each \(n \in \mathbb{N}\).
A 15. The weights \( w_n^i \) for each \( i = 1, \ldots, M_n^I \) and \( n \in \mathbb{N} \) are chosen such that
\[
\left( \frac{\max_{i=1,\ldots,M_n^I} w_n^i}{\min_{i=1,\ldots,M_n^I} w_n^i} \right) \leq K_w^I < \infty \quad \text{for each } n \in \mathbb{N} \quad (99)
\]

With Assumption A 15, suppose we choose our design points such that there exist constants \( 0 < K_y^I, K_z^I < \infty \) where for each \( n \in \mathbb{N} \),
\[
\left\| (\tilde{Y}_n^I)^T \tilde{Y}_n^I \right\|_2 = \lambda_{\max} \left( (\tilde{Y}_n^I)^T \tilde{Y}_n^I \right) < K_y^I < \infty \quad (100)
\]
and
\[
\lambda_{\min} \left( (\tilde{Y}_n^I)^T \tilde{Y}_n^I \right) > K_z^I > 0 \quad \text{i.e.,} \quad \left\| (\tilde{Y}_n^I)^T \tilde{Y}_n^I \right\|_2 < \frac{1}{K_z^I} < \infty \quad (101)
\]
Then, it is clear from (97) that (36) is satisfied for \( K^I = K_w^I \left( K_y^I / K_z^I \right) \). Analogously, suppose we choose our design points such that there exist some constants \( 0 < K_z^I < \infty \) where for each \( n \in \mathbb{N} \),
\[
\left\| (\tilde{Z}_n^I)^T \tilde{Z}_n^I \right\|_2 = \lambda_{\max} \left( (\tilde{Z}_n^I)^T \tilde{Z}_n^I \right) < K_z^I < \infty \quad (102)
\]
and
\[
\lambda_{\min} \left( (\tilde{Z}_n^I)^T \tilde{Z}_n^I \right) > K_z^I > 0 \quad \text{i.e.,} \quad \left\| (\tilde{Z}_n^I)^T \tilde{Z}_n^I \right\|_2 < \frac{1}{K_z^I} < \infty \quad (103)
\]
Then, from Assumption A 15 and (98), we get that (58) is satisfied for \( K^I = K_w^I \left( K_z^I / K_z^I \right) \). Accordingly, in this section, we will provide methods to pick the design points such that (100), (101), (102) and (103) hold for each \( n \in \mathbb{N} \). In Section 3.2.3, we will provide an example of a scheme to set the weights \( \{w_n^i : i = 1, \ldots, M_n^I \} \) such that (99) holds for some \( K_w^I < \infty \) and each \( n \in \mathbb{N} \).

First, let us consider (100) and (102). We know that by definition, the inner design points satisfy
\[
\| \tilde{y}_n^i \|_2 \leq \delta_n, \text{ i.e., } \| \tilde{y}_n^i \|_2 \leq 1 \quad \text{for each } i = 1, \ldots, M_n^I.
\]
Using (17) we then get that \( \| \tilde{z}_n^i \|_2 \leq (\sqrt{3}/2) \) for each \( i = 1, \ldots, M_n^I \). Now, if we can ensure that for each \( n \in \mathbb{N} \), the design points we pick satisfy the following assumption

A 16. For each \( n \in \mathbb{N} \), \( M_n^I \leq M_{\text{max}}^I \) for some constant \( M_{\text{max}}^I < \infty \).

Then from (95) we get that for each \( n \in \mathbb{N} \),
\[
\left\| (\tilde{Y}_n^I)^T \tilde{Y}_n^I \right\|_2 = \max_{x \in \mathbb{R}^\prime, \|x\|_2 = 1} \sum_{i=1}^{M_n^I} (\tilde{y}_n^i)^T x)^2 \leq M_n^I \max_{i \in \{1, \ldots, M_n^I\}} \| \tilde{y}_n^i \|_2^2 \leq M_{\text{max}}^I \quad (104)
\]
\[
\left\| (\tilde{Z}_n^I)^T \tilde{Z}_n^I \right\|_2 = \max_{x \in \mathbb{R}^\prime, \|x\|_2 = 1} \sum_{i=1}^{M_n^I} (\tilde{z}_n^i)^T x)^2 \leq M_n^I \max_{i \in \{1, \ldots, M_n^I\}} \| \tilde{z}_n^i \|_2^2 \leq \left( \frac{3}{2} \right) M_{\text{max}}^I \quad (105)
\]

Next, we consider (101) and (103). In order to see that these are conditions on the relative positions of the inner design points within the design region, we first state some well-known facts in basic linear algebra and set the related notation.
Projection Operation: For any \( y \in \mathbb{R}^l \) and any subspace \( \mathcal{M} \subset \mathbb{R}^l \), we let

\[
\Pi(y, \mathcal{M}) := \arg \min_{x \in \mathcal{M}} \|y - x\|_2
\]

(106)
denote the orthogonal projection of \( y \) onto \( \mathcal{M} \). The following properties of the projection operator are well known.

**P 3.6.** For any \( y \in \mathbb{R}^l \) and subspace \( \mathcal{M} \subset \mathbb{R}^l \), we have \( \|y\|_2 \geq \|\Pi(y, \mathcal{M})\|_2 \).

**P 3.7.** For any \( c_1, c_2 \in \mathbb{R}, y^1, y^2 \in \mathbb{R}^l \) and subspace \( \mathcal{M} \subset \mathbb{R}^l \), we have

\[
\Pi(c_1 y^1 + c_2 y^2, \mathcal{M}) = c_1 \Pi(y^1, \mathcal{M}) + c_2 \Pi(y^2, \mathcal{M})
\]

**P 3.8.** Given a subspace \( \mathcal{M} \subset \mathbb{R}^l \) is a subspace with \( \dim(\mathcal{M}) = j \in \{1, \ldots, l\} \) and an orthonormal basis \( \{v^1, \ldots, v^j\} \) for \( \mathcal{M} \), we have

\[
\Pi(y, \mathcal{M}) = \sum_{i=1}^{j} (v^i y^i) v^i
\]

for any \( y \in \mathbb{R}^l \). Consequently, \( \|\Pi(y, \mathcal{M})\|_2 = \sqrt{\sum_{i=1}^{j} (y^i v^i)^2} \).

Orthogonal Complements: Given any subspace \( \mathcal{M} \subset \mathbb{R}^l \), we will let \( \mathcal{M}^\perp \) denote the orthogonal complement of \( \mathcal{M} \) in \( \mathbb{R}^l \).

**P 3.9.** For any \( y \in \mathbb{R}^l \) and subspace \( \mathcal{M} \subset \mathbb{R}^l \), there exist unique vectors \( y^\mathcal{M} \in \mathcal{M} \) and \( y^\mathcal{M}^\perp \in \mathcal{M}^\perp \) such that \( y = y^\mathcal{M} + y^\mathcal{M}^\perp \). In fact, it can be shown that \( y^\mathcal{M} = \Pi(y, \mathcal{M}) \) and \( y^\mathcal{M}^\perp = \Pi(y, \mathcal{M}^\perp) \).

Consequently, using (106), we get the following property.

**P 3.10.** For any \( y \in \mathbb{R}^l \) and subspace \( \mathcal{M} \subset \mathbb{R}^l \), \( \|\Pi(y, \mathcal{M}^\perp)\|_2 = \|y - \Pi(y, \mathcal{M})\|_2 \) measures the distance of \( y \) from the subspace \( \mathcal{M} \). In particular, \( y \in \mathcal{M} \) if and only if \( y = \Pi(y, \mathcal{M}) \), i.e., if and only if \( \|\Pi(y, \mathcal{M}^\perp)\|_2 = 0 \).

Now, let us again consider the matrix \( Y \in \mathbb{R}^{m \times l} \) defined in (94). The following lemma provides an alternative characterization of the minimum eigenvalue of \( Y^T Y \).

**Lemma 3.11.** Let \( Y \in \mathbb{R}^{m \times l} \) and let \( \{(y^1)^T, \ldots, (y^m)^T\} \subset \mathbb{R}^l \) denote its rows. Then, there exists a subspace \( \mathcal{M}^* \subset \mathbb{R}^l \) with \( \dim(\mathcal{M}) = l - 1 \) such that

\[
\inf_{\substack{\mathcal{M} \subset \mathbb{R}^l \\
\dim(\mathcal{M}) = l - 1}} \sum_{i=1}^{m} \|\Pi(y^i, \mathcal{M}^*)\|_2^2 = \sum_{i=1}^{m} \left\|\Pi\left(y^i, (\mathcal{M}^*)^\perp\right)\right\|_2^2
\]

Further, we have

\[
\lambda_{\text{min}}(Y^T Y) = \min_{\substack{\mathcal{M} \subset \mathbb{R}^l \\
\dim(\mathcal{M}) = l - 1}} \sum_{i=1}^{m} \|\Pi(y^i, \mathcal{M}^*)\|_2^2 = \sum_{i=1}^{m} \left\|\Pi\left(y^i, (\mathcal{M}^*)^\perp\right)\right\|_2^2
\]

(107)

Consequently, \( \lambda_{\text{min}}(Y^T Y) = 0 \) if and only if the row rank of \( Y \) is less than \( l \).
Proof. We know that for each subspace \( M \subset \mathbb{R}^l \) with \( \dim(M) = l - 1 \), \( \dim(M^\perp) = 1 \) and consequently, there exists \( x \in M^\perp \subset \mathbb{R}^l \) with \( \|x\|_2 = 1 \) such that \( M^\perp = \text{span}\{x\} \). Similarly, for each \( x \in \mathbb{R}^l \) with \( \|x\|_2 = 1 \), \( \text{span}\{x\} \) is a subspace of dimension 1 and consequently, the subspace \( M = (\text{span}\{x\})^\perp \) satisfies \( \dim(M) = l - 1 \).

Finally, using (96), it is clear that (107) holds. Therefore, setting

\[
\Pi(y^i, M^\perp) = (y^i)^T x \quad \text{and} \quad \|\Pi(y^i, M^\perp)\|_2 = (y^i)^T x
\]

and consequently

\[
\sum_{i=1}^m \|\Pi(y^i, M^\perp)\|_2^2 = \sum_{i=1}^m (y^i)^T x^2
\]

(108)

Therefore, we clearly get that

\[
\inf_{M \subset \mathbb{R}^l, \dim(M) = l-1} \sum_{i=1}^m \|\Pi(y^i, M^\perp)\|_2^2 = \inf_{x \in \mathbb{R}^l, \|x\|_2 = 1} \sum_{i=1}^m (y^i)^T x^2
\]

(109)

Now, since \( \sum_{i=1}^m (y^i)^T x^2 \) is a continuous function of \( x \) and \( \{x \in \mathbb{R}^l : \|x\|_2 = 1\} \) is a compact set, we know that there exists \( x^* \in \mathbb{R}^l \) with \( \|x^*\|_2 = 1 \) such that

\[
\sum_{i=1}^m (y^i)^T x^* = \inf_{x \in \mathbb{R}^l, \|x\|_2 = 1} \sum_{i=1}^m (y^i)^T x^2
\]

Therefore, setting \( M^* = (\text{span}\{x^*\})^\perp \), and using (108) and (109), we get

\[
\sum_{i=1}^m \|\Pi(y^i, (M^*)^\perp)\|_2^2 = \sum_{i=1}^m (y^i)^T x^* = \inf_{x \in \mathbb{R}^l, \|x\|_2 = 1} \sum_{i=1}^m (y^i)^T x^2
\]

Finally, using (96), it is clear that (107) holds.

In order to show the final assertion in Lemma 3.11, let us first assume that \( \lambda_{\text{min}}(Y^T Y) = 0 \). Then, from (107), there exists a subspace \( M^* \subset \mathbb{R}^l \) with \( \dim((M^*)^\perp) = l - 1 \), such that

\[
\sum_{i=1}^m \|\Pi(y^i, (M^*)^\perp)\|_2^2 = 0
\]

This means that \( \Pi(y^i, (M^*)^\perp) = 0 \) for each \( i = 1, \ldots, m \). Thus using Property P 3.10, we get that \( y^i \in M^* \) for each \( i = 1, \ldots, m \). But now, since \( \text{span}\{y^1, \ldots, y^m\} \subset M^* \) and \( \dim(M^*) = l - 1 \), we get that the row rank of \( Y^T Y \) can be at most \( l - 1 \).

Conversely, if the row rank of \( Y^T Y \) is less than \( l \), then \( \dim(\text{span}\{y^1, \ldots, y^m\}) < l \). Therefore, for \( x^* \in (\text{span}\{y^1, \ldots, y^m\})^\perp \) such that \( \|x^*\|_2 = 1 \), we get

\[
\sum_{i=1}^m (y^i)^T x^* = 0
\]
Therefore, using (96), we get that

\[ 0 \leq \lambda_{\min}(Y^T Y) \leq \sum_{i=1}^{m} (y_i^T x^*)^2 = 0 \]

\[ \square \]

Using Property P 3.10 and Lemma 3.11, we can state (107) as follows. The minimum eigenvalue of the matrix \( Y^T Y \) is equal to minimum of the sum of squared distances of the rows vectors of \( Y \) to any subspace \( \mathcal{M} \subset \mathbb{R}^l \) of dimension equal to \( l - 1 \). Thus, it is intuitively clear that \( \lambda_{\min}(Y^T Y) \) is close to zero whenever there exists a subspace \( \mathcal{M} \subset \mathbb{R}^l \) of dimension \( l - 1 \) such that \( \|\Pi(y, \mathcal{M})\|_2 \) is small for each \( i = 1, \ldots, m \), i.e., all the rows vectors of \( Y \) lie close to \( \mathcal{M} \). Indeed as we showed in Lemma 3.11, we get \( \lambda_{\min}(Y^T Y) = 0 \) whenever the distances from each of the rows of \( Y \) to some subspace is equal to zero, i.e., the rank of \( Y \) is less than \( l \).

In order to formalize this intuitive notion, we present next, a series of results quantitatively relating the geometry of the vectors \( \{y^1, \ldots, y^m\} \) to the minimum eigenvalue of \( Y^T Y \). First, let us consider the vectors corresponding to the first \( l \) rows \( \{y^1, \ldots, y^l\} \subset \mathbb{R}^l \). For the sake of notational brevity and consistency we define \( Q_i := \text{span}\{y^1, \ldots, y^i\} \) for each \( i = 1, \ldots, l \). For the present, we will assume that for each \( i = 1, \ldots, l \),

\[ \|y^i\|_2 = 1 \quad (110) \]

and that there exists \( \mu \in (0, 1] \) such that for \( i = 2, \ldots, l \)

\[ \|\Pi(y^i, (Q^{i-1})^\perp)\|_2 \geq \mu \quad (111) \]

Note that from (110), \( \|y^1\|_2 > 0 \). Also, for each \( i = 2, \ldots, l \), (111) implies that the distance \( \|\Pi(y^i, Q^{i-1})\|_2 \) from \( y^i \) to \( \text{span}\{y^1, \ldots, y^{i-1}\} \) has to be at least \( \mu > 0 \). In particular, this means that \( y^i \notin \text{span}\{y^1, \ldots, y^{i-1}\} \).

Naturally, under these conditions, we would expect the set \( \{y^1, \ldots, y^l\} \) to be linearly independent. The following lemma shows this result.

**Lemma 3.12.** Consider \( j \leq l \) vectors \( x^1, \ldots, x^j \in \mathbb{R}^l \) that satisfy

\[ \|x^1\|_2 > 0 \quad \text{and if} \; j \geq 2, \quad \|\Pi(x^i, \text{span}\{x^1, \ldots, x^{i-1}\}^\perp)\|_2 > 0 \quad \text{for each} \; i = 2, \ldots, j \quad (112) \]

Then, the vectors \( \{x^1, \ldots, x^j\} \) are linearly independent.

**Proof.** If \( j = 1 \), then it is trivially clear that \( \{x^1\} \) is a linearly independent set since \( \|x^1\|_2 > 0 \). Therefore, let us assume that \( j \geq 2 \).

First, using Property P 3.6 it is easily seen that (112) implies \( \|x^i\|_2 > 0 \) for each \( i = 1, \ldots, j \). Further, it is also clear from Property P 3.10 that for each \( i = 2, \ldots, j \),

\[ x^i \notin \text{span}\{x^1, \ldots, x^{i-1}\} \quad (113) \]

53
Now, in order to show that the vectors \( \{x^1, \ldots, x^j\} \) are linearly independent, we must show that the only solution to
\[
\alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_j x^j = 0
\]
is \( \alpha_1 = \alpha_2 = \ldots = \alpha_j = 0 \). We show this by contradiction.

Let us assume that there exists a solution to (114) with \( \alpha_j \neq 0 \). In this case, we can write,
\[
x^j = \left(\frac{-1}{\alpha_j}\right) (\alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_{j-1} x^{j-1})
\]
If \( \alpha_1 = \ldots = \alpha_{j-1} = 0 \), then we get that \( x^j = 0 \) which contradicts the fact that \( \|x^j\|_2 > 0 \). Therefore, there exist some \( k \in \{1, \ldots, j-1\} \) such that \( \alpha_k \neq 0 \). Now this means that \( x^j \in \text{span}\{x^1, \ldots, x^{j-1}\} \). This contradicts (113) for \( i = j \). Therefore, there exists no solution for (114) with \( \alpha_j \neq 0 \).

Now assume that there exists a solution to (114) with \( \alpha_{j-1} \neq 0 \). Then, since \( \alpha_j = 0 \), we can write
\[
x^{j-1} = \left(\frac{1}{\alpha_{j-1}}\right) (\alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_{j-2} x^{j-2})
\]
As before, since \( \|x^{j-1}\|_2 > 0 \), there exists some \( j \in \{1, \ldots, j-2\} \), such that \( \alpha_k \neq 0 \). This means that \( x^{j-1} \in \text{span}\{x^1, \ldots, x^{j-2}\} \), i.e., which contradicts (113) for \( i = j - 1 \). Thus, \( \alpha_{j-1} = 0 \).

Proceeding in the same manner as above, we can show that \( \alpha_2 = \ldots = \alpha_j = 0 \). But now, since \( \|x^1\|_2 > 0 \), we get that the only solution to \( \alpha_1 x^1 = 0 \) is \( \alpha_1 = 0 \).

Thus, we have shown that the only solution to (114) is \( \alpha_1 = \ldots = \alpha_j = 0 \). Therefore, the vectors \( \{x^1, \ldots, x^j\} \) are linearly independent.

Now, let us apply Lemma 3.12 to the vectors \( \{y^1, \ldots, y^l\} \subset \mathbb{R}^l \) that satisfy (110) and (111) for each \( i = 1, \ldots, l \). From Property P 3.6, it is clear that the conditions (111) for \( i = 1, \ldots, l \) together satisfy (112). Therefore we get that the vectors \( \{y^1, \ldots, y^l\} \) are linearly independent. Consequently, the row rank of \( Y \) (defined in (94)), is equal to \( l \) and from Lemma 3.11, we get \( \lambda^\text{min}(Y^T Y) > 0 \). However, we want to show not only that \( \lambda^\text{min}(Y^T Y) > 0 \) but also that there exists a positive lower bound on this minimum eigenvalue that is a function of \( \mu \). To this end, we will first show that if for each \( i = 1, \ldots, l \), (110) is satisfied and (111) is satisfied for some \( \mu \in (0, 1] \), then we get
\[
\min_{x \in \mathbb{R}^l} \frac{1}{\|x\|_2} \sum_{i=1}^l ((y^i)^T x)^2 \geq S_l(\mu)
\]
where \( S_l(\mu) \) is the \( l \)-th term in the sequence \( \{S_k(\mu)\}_{k \in \mathbb{N}} \subset \mathbb{R} \) defined iteratively for any \( \mu \in (0, 1] \), as follows.
\[
S_1(\mu) := 1 \quad \text{and} \quad S_j(\mu) = \left(1 + S_{j-1}(\mu)\right) \left[ 1 - \sqrt{1 - \frac{4S_{j-1}(\mu) \mu^2}{(S_{j-1}(\mu) + 1)^2}} \right] \quad \text{for} \quad j = 2, 3, \ldots
\]
The following lemma shows that the sequence \( \{S_j(\mu)\}_{j \in \mathbb{N}} \) is well defined for any \( \mu \in (0, 1] \) and that \( S_j(\mu) > 0 \) for each \( j \in \mathbb{N} \).
Lemma 3.13. Consider the sequence \( \{S_j(\mu)\}_{j \in \mathbb{N}} \subset \mathbb{R} \) defined in (116). For any \( \mu \in (0, 1] \), the following assertions hold.

1. For each \( j \in \mathbb{N} \), \( S_j(\mu) \) is a well defined real number and in particular \( S_j(\mu) > 0 \).

2. For each \( j \geq 2 \), \( S_j(\mu) \leq \mu^2 S_{j-1}(\mu) \).

Proof. We show the first assertion by induction. First, it is trivially clear that \( S_1(\mu) \) is a real number and \( S_1(\mu) > 0 \). Next, let us suppose that \( S_{j-1}(\mu) \in \mathbb{R} \) and \( S_{j-1}(\mu) > 0 \) for some \( j \geq 2 \). Let us rearrange the expression for \( S_j(\mu) \) (for \( j \geq 2 \)) in (116) as follows.

\[
S_j(\mu) = \left( \frac{1 + S_{j-1}(\mu)}{2} \right) - \left( \frac{1}{2} \right) \sqrt{(1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2} \quad (117)
\]

Then, since \( \mu \in (0, 1] \),

\[
(1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2 \geq (1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu) = (1 - S_{j-1}(\mu))^2 \geq 0 \quad (118)
\]

Thus, \( \sqrt{(1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2} \in \mathbb{R} \) and \( S_j(\mu) \) is a well-defined real number. Further, it is easily seen that since \( S_{j-1}(\mu) > 0 \) and \( \mu > 0 \), we have \( 4S_{j-1}(\mu)\mu^2 > 0 \). Therefore,

\[
\sqrt{(1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2} < (1 + S_{j-1}(\mu))
\]

\[
\Rightarrow \quad (1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2 < (1 + S_{j-1}(\mu))^2
\]

\[
\Rightarrow \quad (1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2 < (1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2
\]

\[
\Rightarrow \quad \left( \frac{1 + S_{j-1}(\mu)}{2} \right) - \left( \frac{1}{2} \right) \sqrt{(1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2} > 0
\]

\[
\Rightarrow \quad S_j(\mu) > 0
\]

Therefore we get that \( S_j(\mu) > 0 \). Consequently, by induction we get that \( S_j(\mu) \in \mathbb{R} \) and \( S_j(\mu) > 0 \) for each \( j \in \mathbb{N} \).

Next let us consider the second assertion. Since \( \mu \in (0, 1] \) and \( S_j(\mu) > 0 \) for each \( j \in \mathbb{N} \), we know that \( 4S_{j-1}(\mu)\mu^2(\mu^2 - 1) \leq 0 \) for each \( j \geq 2 \). Therefore, we get for each \( j \geq 2 \),

\[
\frac{(1 + S_{j-1}(\mu))^2 + 4S_{j-1}(\mu)\mu^2(\mu^2 - 1) - 4S_{j-1}(\mu)\mu^2}{4} \leq \frac{(1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2}{4}
\]

\[
\Rightarrow \quad \left( \frac{1 + S_{j-1}(\mu) - 2S_{j-1}(\mu)\mu^2}{2} \right)^2 \leq \left( \frac{1}{2} \right) \sqrt{(1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2}^2
\]

From (118), we know that the square root on the right side of the final inequality given above exists (i.e., is a real number). Now, for any \( a, b \in \mathbb{R} \), we know that if \( a^2 \leq b^2 \), then \( a \leq |a| \leq |b| \). Therefore, we get that for each \( j \geq 2 \),

\[
\frac{1 + S_{j-1}(\mu) - 2S_{j-1}(\mu)\mu^2}{2} \leq \frac{1}{2} \sqrt{(1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2}
\]

\[
\Rightarrow \quad \left( \frac{1 + S_{j-1}(\mu)}{2} \right) - \left( \frac{1}{2} \right) \sqrt{(1 + S_{j-1}(\mu))^2 - 4S_{j-1}(\mu)\mu^2} \leq S_{j-1}(\mu)\mu^2
\]

\[
\Rightarrow \quad S_j(\mu) \leq S_{j-1}(\mu)\mu^2
\]
Now, in order to show that (115) satisfied, we will use the Gram-Schmidt orthogonalization process on the vectors \( \{y^1, \ldots, y^l\} \). Accordingly, before we do so, let us review some well known properties of the Gram-Schmidt Process.

**Gram-Schmidt Orthogonalization:** Consider any \( 1 \leq j \leq l \) linearly independent vectors \( \{x^1, \ldots, x^j\} \subset \mathbb{R}^l \). We can successfully perform the Gram-Schmidt orthogonalization process on this set. That is, we can define the sets of vectors \( \{u^1, \ldots, u^j\} \subset \mathbb{R}^l \) and \( \{q^1, \ldots, q^j\} \subset \mathbb{R}^l \) as follows.

\[
\begin{align*}
  u^1 & := x^1 \quad \text{and} \quad q^1 := \frac{u^1}{\|u^1\|_2} \\
  u^i & := x^i - \sum_{k=1}^{i-1} \left( x^iT q^k \right) q^k \quad \text{and} \quad q^i := \frac{u^i}{\|u^i\|_2} \quad \text{for each} \quad i = 2, \ldots, j
\end{align*}
\]  

(119)

and if \( j \geq 2 \) then

\[
\begin{align*}
  u^i & := x^i - \sum_{k=1}^{i-1} \left( x^iT q^k \right) q^k \quad \text{and} \quad q^i := \frac{u^i}{\|u^i\|_2} \quad \text{for each} \quad i = 2, \ldots, j
\end{align*}
\]  

(120)

The following properties of \( \{u^1, \ldots, u^j\} \) and \( \{q^1, \ldots, q^j\} \) are well-known.

**P 3.11.** \( \{u^1, \ldots, u^j\} \) is an orthogonal set of vectors and \( \{q^1, \ldots, q^j\} \) is an orthonormal set of vectors.

**P 3.12.**

1. For each \( i = 1, \ldots, j \),

\[
\text{span}\{x^1, \ldots, x^i\} = \text{span}\{u^1, \ldots, u^i\} = \text{span}\{q^1, \ldots, q^i\}
\]

2. Thus, from Property P 3.11, we get that for each \( i = 1, \ldots, j \), \( \{q^1, \ldots, q^i\} \) forms an orthonormal basis for \( \text{span}\{x^1, \ldots, x^i\} \). In particular, if \( j = l \), then \( \{q^1, \ldots, q^l\} \) forms an orthonormal basis for \( \text{span}\{x^1, \ldots, x^l\} = \mathbb{R}^l \).

**P 3.13.** If \( 2 \leq j \leq l \), then for each \( i = 2, \ldots, j \), using Property P 3.8 we get,

\[
\Pi \left( x^i, \text{span}\{x^1, \ldots, x^{i-1}\} \right) = \sum_{k=1}^{i-1} \left( x^iT q^k \right) q^k
\]

But we know from (123) that for each \( i = 2, \ldots, j \), \( x^i = \sum_{k=1}^{i} \left( x^iT q^k \right) q^k \). Therefore, using the fact that

\[
x^i = \Pi \left( x^i, \text{span}\{x^1, \ldots, x^{i-1}\} \right) + \Pi \left( x^i, (\text{span}\{x^1, \ldots, x^{i-1}\})^\perp \right),
\]

we get

\[
\Pi \left( x^i, (\text{span}\{x^1, \ldots, x^{i-1}\})^\perp \right) = x^i - \sum_{k=1}^{i-1} \left( x^iT q^k \right) q^k = u^i = \left( x^iT q^i \right) q^i
\]

Consequently, we get that if \( j \geq 2 \), then for each \( i = 2, \ldots, j \), \( \text{span}\{q^i\} \) is the orthogonal complement of the subspace \( Q^{i-1} \) of the vector space \( Q^i \). That is, any \( x \in Q^i \) can be written uniquely as the sum of two vectors \( \Pi(x, Q^{i-1}) \in Q^{i-1} \) and \( (x^T q^i) q^i \in \text{span}\{q^i\} \).
Now, analogous to (119) and (120), we define the Grams-Schmidt vectors corresponding to the set \( \{y^1, \ldots, y^l\} \subset \mathbb{R}^l \) (satisfying (110) and (111)) as
\[
\begin{align*}
  u_1 &:= y^1 \\
  q_1 &:= \frac{u_1}{\|u_1\|_2} \\
  u_i &:= y^i - \sum_{k=1}^{i-1} (y^i)^T q^k q^k \\
  q_i &:= \frac{u_i}{\|u_i\|_2}
\end{align*}
\]
for each \( i = 2, \ldots, l \) (121)

From Property P 3.12 we know that for \( i = 1, \ldots, l \), \( \{q^1, \ldots, q^i\} \) forms an orthonormal basis for \( Q^i \). Thus, for any \( x \in Q^i \), we can write
\[
  x = \sum_{k=1}^{i} (x^T q^k) q^k \\
  \|x\|_2 = \sum_{k=1}^{i} (x^T q^k)^2
\]
where \( (x^T q^1, \ldots, x^T q^i) \) represent the unique coordinates of \( x \) in \( Q^i \), with respect to the orthonormal basis \( \{q^1, \ldots, q^i\} \). Since we will repeatedly use such coordinates in our analysis, for the sake of notational brevity, we let \( x^q_i = x^T q^i \) denote the coordinate of any vector \( x \in \mathbb{R}^l \), in the direction of the basis vector \( q^i \) for any \( i = 1, \ldots, l \).

Since we intend to show using induction that (115) is satisfied, we start with a result involving only the first two vectors \( y^1 \) and \( y^2 \) from our set \( \{y^1, \ldots, y^l\} \).

**Lemma 3.14.** Consider the vectors \( y^1, y^2 \in \mathbb{R}^l \) from the set \( \{y^1, \ldots, y^l\} \in \mathbb{R}^l \) that satisfies (110) and (111). We have,
\[
\min_{x \in Q^2, \|x\|_2 = 1} \left( y^1^T x \right)^2 + \left( y^2^T x \right)^2 \geq S_2(\mu)
\]
where \( S_2(\mu) \) is the second term in the sequence \( \{S_j(\mu)\}_{j \in \mathbb{N}} \) defined in (116). Consequently,
\[
\left( y^1^T x \right)^2 + \left( y^2^T x \right)^2 \geq S_2(\mu) \|x\|_2^2
\]
for all \( x \in Q^2 \).

In order to prove Lemma 3.14, we will need the following lemma regarding a related one dimensional optimization problem. In what follows and for the rest of this article we assume that for any non-negative real number \( \alpha \), \( \sqrt{\alpha} \) denotes the positive square root of \( \alpha \). In particular, for any \( \alpha \in \mathbb{R} \), \( \sqrt{\alpha^2} = |\alpha| \).

**Lemma 3.15.** Let \( s, a \in (0, 1] \). Consider the function \( g : [0, 1] \rightarrow \mathbb{R} \) defined as
\[
g(t) := \left( at - \sqrt{1-a^2} \sqrt{1-t^2} \right)^2 + s \left( 1 - t^2 \right)
\]
Then,
\[
\min_{t \in [0, 1]} g(t) = \left( \frac{1+s}{2} \right) \left( 1 - \sqrt{1 - \frac{4sa^2}{(1+s)^2}} \right)
\]
Consequently, for any \( \mu \in (0, a] \), we get
\[
\min_{t \in [0, 1]} g(t) \geq \left( \frac{1+s}{2} \right) \left( 1 - \sqrt{1 - \frac{4s\mu^2}{(1+s)^2}} \right)
\]
Proof. First, we note that for any \( s, a \in \mathbb{R} \),

\[
(1 + s)^2 - 4sa^2 = (1 + s - 2a^2)^2 + 4a^2(1 - a^2)
\]

(128)

Consequently, for any \( a \in (0, 1] \), we get that \((1 + s)^2 - 4sa^2 \geq 0\). Therefore, the minimum objective value given on the right side of (126) is a real number and hence the statement of Lemma 3.15 is well defined.

Now, we break the assertion in (126) into two cases.

1. First, let us consider the case when \( a = 1 \). Then, it can be seen from (125) that \( g(t) = s + (1 - s)t^2 \).

Thus, since \( 1 - s \geq 0 \), obviously, \( \min_{t \in [0, 1]} g(t) = s \). But now, using the fact that \( s \in (0, 1] \), we get

\[
\left( \frac{1 + s}{2} \right) \left( 1 - \sqrt{1 - \frac{4s}{(1 + s)^2}} \right) = \left( \frac{1 + s}{2} \right) \left( 1 - \frac{1 - s}{1 + s} \right)
\]

\[
= \left( \frac{1 + s}{2} \right) \left( 1 - \frac{1 - s}{1 + s} \right) = s
\]

Therefore, (126) holds for \( a = 1 \).

2. Next, we assume that \( a \in (0, 1) \). In this case, note that we get from (128) that

\[
(1 + s)^2 - 4sa^2 > 0
\]

(129)

Now, the function \( g \) can be rearranged and written as

\[
g(t) = \left( at - \sqrt{1 - a^2} \sqrt{1 - t^2} \right)^2 + s(1 - t^2)
\]

\[
= a^2t^2 + (1 - a^2)(1 - t^2) - 2(a\sqrt{1 - a^2})(t\sqrt{1 - t^2}) + s(1 - t^2)
\]

\[
= (2a^2 - s - 1)t^2 - 2(a\sqrt{1 - a^2})(t\sqrt{1 - t^2}) + (1 + s - a^2)
\]

It is easily seen that \( g \) is differentiable on \((0, 1)\). Accordingly, let us find the values of \( t \) in the interval \((0, 1)\) that satisfy \( \frac{dg}{dt}(t) = 0 \). We have,

\[
\frac{dg}{dt}(t) = 2(2a^2 - s - 1)t - 2(a\sqrt{1 - a^2})(\sqrt{1 - t^2}) + \frac{2(a\sqrt{1 - a^2})t^2}{\sqrt{1 - t^2}}
\]

\[
= 2 \left[ (2a^2 - s - 1)(t\sqrt{1 - t^2}) - (a\sqrt{1 - a^2})(1 - 2t^2) \right]
\]

Thus, in order to find first order critical points of \( g \) in \((0, 1)\), we must find the solutions in \((0, 1)\) to the equation

\[
(2a^2 - s - 1)(t\sqrt{1 - t^2}) = (a\sqrt{1 - a^2})(1 - 2t^2)
\]

(130)

Squaring both sides to remove the square roots and rearranging, we get the following quadratic equation in \( t^2 \).

\[
((s + 1)^2 - 4sa^2)t^4 - ((s + 1)^2 - 4sa^2)t^2 + a^2(1 - a^2) = 0
\]

(131)
Let us set \( A = (s + 1)^2 - 4a^2 \), \( B = -A \) and \( C = a^2(1 - a^2) \). From (129), we get that \( A > 0 \), \( B < 0 \) and \( C > 0 \). Then the quadratic equation in (131) is given by \( At^4 + Bt^2 + C = 0 \). Now, we get

\[
B^2 - 4AC = ((s + 1)^2 - 4a^2)^2 - 4((s + 1)^2 - 4a^2)a^2(1 - a^2)
= (s + 1)^2 - 4a^2(s + 2a^2) ^2 \geq 0
\]

Therefore, we get two real-valued solutions that satisfy (131), given by

\[
t^2 = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} = \left\{ \begin{array}{ll}
\left(\frac{1}{2}\right) \left\{ 1 + \frac{s + 1 - 2a^2}{\sqrt{(s + 1)^2 - 4a^2}} \right\} \\
\left(\frac{1}{2}\right) \left\{ 1 - \frac{s + 1 - 2a^2}{\sqrt{(s + 1)^2 - 4a^2}} \right\}
\end{array} \right.
\]

Now, since we have assumed that \( a \in (0, 1) \), we get from (128) that

\[
\left| \frac{s + 1 - 2a^2}{\sqrt{(s + 1)^2 - 4a^2}} \right| < 1
\] (132)

Thus, we get the following four potential real-valued solutions to (130)

\[
t_1^* = \sqrt{\left(\frac{1}{2}\right) \left\{ 1 + \frac{s + 1 - 2a^2}{\sqrt{(s + 1)^2 - 4a^2}} \right\}} \quad \text{and} \quad t_2^* = \sqrt{\left(\frac{1}{2}\right) \left\{ 1 - \frac{s + 1 - 2a^2}{\sqrt{(s + 1)^2 - 4a^2}} \right\}}
\]

\[
t_3^* = -\sqrt{\left(\frac{1}{2}\right) \left\{ 1 + \frac{s + 1 - 2a^2}{\sqrt{(s + 1)^2 - 4a^2}} \right\}} \quad \text{and} \quad t_4^* = -\sqrt{\left(\frac{1}{2}\right) \left\{ 1 - \frac{s + 1 - 2a^2}{\sqrt{(s + 1)^2 - 4a^2}} \right\}}
\]

Since we are interested only in solutions in the interval \([0, 1]\), we consider only the two solutions \( t_1^* \) and \( t_2^* \). From (132) it is easily seen that \( t_1^*, t_2^* \in (0, 1) \).

Now, substituting \( t_1^* \) and \( t_2^* \) back into (130), it is easily seen that only \( t_1^* \) satisfies (130). Therefore, there is exactly one point \( t_1^* \in (0, 1) \) such that \( \frac{dg}{dt}(t_1^*) = 0 \). Consequently, we get that

\[
\min_{t \in [0, 1]} g(t) = \min\{g(0), g(1), g(t_1^*)\}
\]

It is easy to verify that

\[
g(0) = 1 + s - a^2 \\
g(1) = a^2 \quad \text{and} \\
g(t_1^*) = \left(\frac{1 + s}{2}\right) \left( 1 - \sqrt{1 - \frac{4sa^2}{(1 + s)^2}} \right)
\]

Further, using (132) and the triangle inequality, we get that

\[
g(1) - g(t_1^*) = \left(\frac{1}{2}\right) \left( \sqrt{(s + 1)^2 - 4sa^2} + (2a^2 - 1 - s) \right) \geq \left(\frac{1}{2}\right) \left( \sqrt{(s + 1)^2 - 4sa^2} - |2a^2 - 1 - s| \right) \geq 0
\]

59
Similarly,
\[ g(0) - g(t_1^*) = \left( \frac{1}{2} \right) \left( \sqrt{(s + 1)^2 - 4sa^2 + (1 + s - 2a^2)} \right) \geq \left( \frac{1}{2} \right) \left( \sqrt{(s + 1)^2 - 4sa^2 - |1 + s - 2a^2|} \right) \geq 0 \]

Therefore, we finally get that (126) holds for the case when \( a \in (0, 1) \).

Thus, from the two cases considered above, we get that (126) holds for any \( s, a \in (0, 1) \).

Now for any \( \mu \in (0, a] \), it is easily seen that
\[
\left( \frac{1 + s}{2} \right) \left( 1 - \sqrt{1 - \frac{4sa^2}{(1 + s)^2}} \right) \geq \left( \frac{1 + s}{2} \right) \left( 1 - \sqrt{1 - \frac{4s\mu^2}{(1 + s)^2}} \right)
\]

Therefore (127) holds for all \( \mu \in (0, a] \).

\[\square\]

**Proof of Lemma 3.14:**

Using the notation that we developed earlier, we write the vectors \( y^1 \in Q^1 \) and \( y^2 \in Q^2 \) in terms of the orthonormal vectors \( q^1 \) and \( q^2 \), as follows. First, since \( y^1 \in Q^1 \) and \( q^1 = (y^1 / \|y^1\|_2) \) (from (122)) is a basis for \( Q^1 \), we write that \( y^1 = y^1_{q^1} q^1 \) where \( y^1_{q^1} = \|y^1\|_2 = 1 \) (using (110)). Similarly, we know from Property P 3.12 that \( \{q^1, q^2\} \) forms an orthonormal basis for \( Q^2 \). Therefore, since \( y^2 \in Q^2 \), we write
\[ y^2 = y^2_{q^1} q^1 + y^2_{q^2} q^2 \]

Here, we get using Property P 3.13 and (111) that
\[ y^2_{q^2} = y^{2T} q^2 = \left\| y^2 - (y^{2T} q^1) q^1 \right\|_2 \geq \mu \]

(133)

Now, from Property P 3.12 we know that \( Q^2 = \{a_1 q^1 + a_2 q^2 : a_1, a_2 \in \mathbb{R} \} \). Therefore, we get that
\[
\{ x \in Q^2 : \|x\|_2 = 1 \} = \{ a_1 q^1 + a_2 q^2 : -1 \leq a_1, a_2 \leq 1 \text{ and } (a_1)^2 + (a_2)^2 = 1 \} \\
= \{ a_1 q^1 + a_2 q^2 : a_2 \in [-1, 1] \text{ and } a_1 = \pm \sqrt{1 - (a_2)^2} \} \\
= \{ a_1 q^1 + a_2 q^2 : |a_2| \leq 1 \text{ and } |a_1| = \sqrt{1 - |a_2|^2} \} \]

(134)

Further, for any \( a_1, a_2 \in \mathbb{R} \) such that \( x = a_1 q^1 + a_2 q^2 \in Q^2 \), we have
\[ y^1 T x = y^1_{q^1} a_1 = a_1 \quad \text{and} \quad y^2 T x = y^2_{q^1} a_1 + y^2_{q^2} a_2 \]

Therefore,
\[ \min_{x \in Q^2, \|x\|_2 = 1} \left( y^1 T x \right)^2 + \left( y^2 T x \right)^2 = \min_{|a_2| \leq 1} \min_{|a_1| = \sqrt{1 - |a_2|^2}} (a_1)^2 + (y^2_{q^1} a_1 + y^2_{q^2} a_2)^2 \]

60
Using the triangle inequality for real numbers, we know that
\[
(y_q^2 a_1 + y_q^2 a_2)^2 \geq (|y_q^2||a_2| - |y_q^2||a_1|)^2
\]
Therefore,
\[
\min_{|a_2| \leq 1} (a_1)^2 + (y_q^2 a_1 + y_q^2 a_2)^2 \geq \min_{|a_1| = \sqrt{1 - |a_2|^2}} (|a_1|)^2 + (|y_q^2||a_2| - |y_q^2||a_1|)^2 \tag{135}
\]
Let us simplify the optimization problem on the right side of (135). First, we know from (110) that \(\|y_q^2\|_2 = 1\). Therefore, we have \((y_q^2)^2 + (y_q^2)^2 = 1\) which gives us
\[
|y_q^2| = \sqrt{1 - |y_q^2|^2} \tag{136}
\]
Similarly, the constraint set in (134) contains the equality constraint \(|a_1| = \sqrt{1 - |a_2|^2}\). We can use this to eliminate the variable \(a_1\) and get
\[
\min_{|a_2| \leq 1} (|a_1|)^2 + (|y_q^2||a_2| - |y_q^2||a_1|)^2
\]
where $S_i(\mu)$ is the $i$th term in the sequence $\{S_j(\mu)\}_{j \in \mathbb{N}}$ defined in (116). Then,

$$\min_{x \in \mathbb{Q}^{i+1}} \sum_{k=1}^{i+1} (y^k T x)^2 \geq S_{i+1}(\mu)$$

Consequently,

$$\sum_{k=1}^{i+1} (y^k T x)^2 \geq S_{i+1}(\mu) \|x\|_2^2 \quad \text{for all } x \in \mathbb{Q}^{i+1}$$

In order to prove Lemma 3.16, we will need the following result.

**Lemma 3.17.** Consider a subspace $M$ of $\mathbb{R}^l$ with $\dim(M) \geq 2$. Further, let $y \in M$ and $a, b > 0$. Then,

$$\min_{x \in M, \|x\|_2 = b} h(x) := (a - |y^T x|)^2 = \begin{cases} 0 & \text{if } a \leq \|y\|_2 b \\ (a - \|y\|_2 b)^2 & \text{if } a > \|y\|_2 b \end{cases}$$

**Proof.** First, we note that since $\dim(M) \geq 2$, the set $S := \{x \in M : \|x\|_2 = b\}$ is a compact and connected set in $M$. Also, the function $\tilde{h}(x) := y^T x$ is continuous on $M$. Therefore, we get that the set $\{y^T x : x \in S\}$ is a connected and compact subset of $\mathbb{R}$; i.e., a closed and bounded interval on $\mathbb{R}$.

Now, from the Cauchy-Schwartz inequality, we know that for any $x \in S$,

$$-\|y\|_2 b = -\|y\|_2 \|x\|_2 \leq y^T x \leq \|y\|_2 \|x\|_2 = \|y\|_2 b$$

Therefore,

$$\{y^T x : x \in S\} \subseteq [-\|y\|_2 b, \|y\|_2 b]$$

Next, consider $x^* = b(y/\|y\|_2) \in M$. Clearly, $\|x\|_2 = b$ and hence $x^* \in S$. Also, it is clear that $y^T x^* = \|y\|_2 b$. Similarly, $-x^* \in S$ and $y^T(-x^*) = -\|y\|_2 b$. Therefore, we get that

$$\{y^T x : x \in S\} \supset [-\|y\|_2 b, \|y\|_2 b]$$

Therefore, we have from (140) and (141) that

$$\{y^T x : x \in S\} = [-\|y\|_2 b, \|y\|_2 b]$$

and hence

$$\{|y^T x| : x \in S\} = [0, \|y\|_2 b]$$

Clearly, $h(x) \geq 0$ for all $x \in S$. Now, if $a \leq \|y\|_2 b$, then we know from (142) that there exists $x^* \in S$ such that $|y^T x^*| = a$; i.e., such that $h(x^*) = 0$. Therefore, if $a \leq \|y\|_2 b$, then

$$\min_{x \in \mathbb{M}, \|x\|_2 = b} (a - |y^T x|)^2 = 0$$
If \( a - \| y \|_2 b > 0 \), then using (142), we get

\[
\{ a - |y^T x| : x \in S \} = [a - \| y \|_2 b, a]
\]

\[
\Rightarrow \{ (a - |y^T x|)^2 : x \in S \} = [(a - \| y \|_2 b)^2, a^2]
\]

Therefore, if \( a - \| y \|_2 b > 0 \), we get

\[
\min_{x \in M, \| x \|_2 = b} (a - |y^T x|)^2 = (a - \| y \|_2 b)^2
\]

\[\square\]

**Proof of Lemma 3.16:**

We know from Property P 3.13 and the notation developed earlier, that any \( x \in Q^{i+1} \) can be written as

\[ x = \Pi(x, Q^i) + (x^T q^{i+1})q^{i+1} = \Pi(x, Q^i) + x_{q^{i+1}}q^{i+1} \]

Conversely, for any \( \bar{x} \in Q^i \subset Q^{i+1} \) and \( a_{i+1} \in \mathbb{R} \), it is obvious that \( \bar{x} + a_{i+1}q^{i+1} \in Q^{i+1} \). Therefore, it is clear that

\[ Q^{i+1} := \{ \bar{x} + a_{i+1}q^{i+1} : \bar{x} \in Q^i \text{ and } a_{i+1} \in \mathbb{R} \} \]

Further, we know from Property P 3.11, that \( q^{i+1} \perp Q^i \). Consequently, we get that for any \( \bar{x} \in Q^i \) and \( a_{i+1} \in \mathbb{R} \), we get that

\[ \| \bar{x} + a_{i+1}q^{i+1} \|_2 = \sqrt{\| \bar{x} \|_2^2 + (a_{i+1})^2} \] (144)

In particular, we get that \( y^{i+1} \in Q^{i+1} \) can be written as

\[ y^{i+1} = \Pi(y^{i+1}, Q^i) + y_{q^{i+1}}^{i+1} \] (145)

Since \( \| y^{i+1} \|_2 = 1 \) (from (110)) we get using (144) that

\[ \| \Pi(y^{i+1}, Q^i) \|_2 = \sqrt{1 - (y_{q^{i+1}}^{i+1})^2} \] (146)

Further, using the fact that \( \| y^{i+1} \|_2 = 1 \) and Property P 3.13 along with (111) for \( y^{i+1} \), we know that

\[ 1 = \| y^{i+1} \|_2 \geq y_{q^{i+1}}^{i+1} = (y^{i+1})^T q^{i+1} = \| \Pi(y^{i+1}, (Q^i)^\perp) \|_2 \geq \mu \] (147)

Also, using (143) and (144), the constraint set in the optimization problem on the left side of (139) can be written as follows.

\[
\{ x \in Q^{i+1} : \| x \|_2 = 1 \} = \left\{ \bar{x} + a_{i+1}q^{i+1} : a_{i+1} \in [-1, 1], \bar{x} \in Q^i \text{ and } \| \bar{x} \|_2 = \sqrt{1 - (a_{i+1})^2} \right\}
\]

\[
= \left\{ \bar{x} + a_{i+1}q^{i+1} : |a_{i+1}| \leq 1, \bar{x} \in Q^i \text{ and } \| \bar{x} \|_2 = \sqrt{1 - |a_{i+1}|^2} \right\}
\]

63
Now, for any \( \bar{x} \in Q^i \) and \( a_{i+1} \in \mathbb{R} \), from (145) and the fact that \( q^{i+1} \perp Q^i \), we get that \( x = \bar{x} + a_{i+1} q^{i+1} \) in \( Q^{i+1} \) satisfies
\[
y^{kT}x = y^{kT}\bar{x} \quad \text{for} \quad k = 1, \ldots, i \quad \text{and} \quad y^{i+1T}x = (\Pi (y^{i+1}, Q^i))^T \bar{x} + y^{i+1}_{q^{i+1}} a_{i+1}.
\]

Therefore, we can now write the optimization problem in (139) as
\[
\min_{x \in Q^{i+1}} \sum_{k=1}^{i+1} (y^{kT}x)^2 = \min_{|a_{i+1}| \leq 1} \left\{ \sum_{k=1}^{i} (y^{kT} x_{1,2})^2 \right\} + \left( (\Pi (y^{i+1}, Q^i))^T \bar{x} + y^{i+1}_{q^{i+1}} a_{i+1} \right)^2
\]

We know from (138) that for any \( \bar{x} \in Q^i \),
\[
\sum_{k=1}^{i} (y^{kT} x_{1,2})^2 \geq S_i(\mu) \| \bar{x} \|^2_2
\]

Also, we know that for any \( \bar{x} + a_{i+1} q^{i+1} \in S_i \), \( \| \bar{x} \|^2_2 = 1 - |a_{i+1}|^2 \). Using these two observations along with the triangle inequality for real numbers we get
\[
\min_{x \in Q^{i+1}} \sum_{k=1}^{i+1} (y^{kT}x)^2 \geq \min_{|a_{i+1}| \leq 1} S_i(\mu) \| \bar{x} \|^2_2 + \left( (\Pi (y^{i+1}, Q^i))^T \bar{x} + y^{i+1}_{q^{i+1}} a_{i+1} \right)^2
\]

Now, let us consider the optimization problem on the right side of (148). We can rewrite this as
\[
\min_{|a_{i+1}| \leq 1} \left[ S_i(\mu) \left( (1 - |a_{i+1}|^2) + \left( |y^{i+1}_{q^{i+1}} | a_{i+1} | - |(\Pi (y^{i+1}, Q^i))^T \bar{x} \right)^2 \right) \right]
\]

For any fixed \( a_{i+1} \in [-1, 1] \) we get by applying Lemma 3.17 and using (146) that
\[
\min_{\| \bar{x} \|^2_2 = \sqrt{1 - |a_{i+1}|^2}} \left( |y^{i+1}_{q^{i+1}} | a_{i+1} | - |(\Pi (y^{i+1}, Q^i))^T \bar{x} \right)^2
\]

It is easy to verify that
\[
|y^{i+1}_{q^{i+1}} | a_{i+1} | \leq \sqrt{1 - |y^{i+1}_{q^{i+1}}|^2} \sqrt{1 - |a_{i+1}|^2} \quad \text{if and only if} \quad |a_{i+1}| \leq \sqrt{1 - |y^{i+1}_{q^{i+1}}|^2}.
\]
Therefore, we finally get
\[
\min_{\tilde{x} \in Q^i, \|\tilde{x}\|_2 = \sqrt{1 - |a_{i+1}|^2}} S_i(\mu) \left( (1 - |a_{i+1}|^2) + \left( |y_{q_{i+1}}^{i+1}| |a_{i+1}| - |(\Pi_{Q^i} y^i)^T \tilde{x}| \right)^2 \right)^{\frac{1}{2}} = \min \{ h_1^*, h_2^* \}
\]
where
\[
h_1^* := \min_{|a_{i+1}| \leq \sqrt{1 - |y_{q_{i+1}}^{i+1}|^2}} S_i(\mu) \left( (1 - |a_{i+1}|^2) \right)
\]
\[
h_2^* := \min_{|a_{i+1}| \in [\sqrt{1 - |y_{q_{i+1}}^{i+1}|^2}, 1]} S_i(\mu) \left( (1 - |a_{i+1}|^2) + \left( |y_{q_{i+1}}^{i+1}| |a_{i+1}| - \sqrt{1 - |y_{q_{i+1}}^{i+1}|^2 \sqrt{1 - |a_{i+1}|^2}} \right)^2 \right)
\]
Using (147), it is easily seen that
\[
h_1^* = S_i(\mu) |y_{q_{i+1}}^{i+1}|^2 \geq S_i(\mu) \mu^2
\]
Also, using Lemma 3.15 with \( p = S_i(\mu) \) and \( a = |y_{q_{i+1}}^{i+1}| \in [\mu, 1] \), we get
\[
h_2^* = \min_{|a_{i+1}| \in [\sqrt{1 - |y_{q_{i+1}}^{i+1}|^2}, 1]} S_i(\mu) \left( (1 - |a_{i+1}|^2) + \left( |y_{q_{i+1}}^{i+1}| |a_{i+1}| - \sqrt{1 - |y_{q_{i+1}}^{i+1}|^2 \sqrt{1 - |a_{i+1}|^2}} \right)^2 \right)
\]
\[
\geq \min_{|a_{i+1}| \in [0, 1]} S_i(\mu) \left( 1 - t^2 \right) + \left( |y_{q_{i+1}}^{i+1}| t - \sqrt{1 - |y_{q_{i+1}}^{i+1}|^2 \sqrt{1 - t^2}} \right)^2
\]
\[
\geq \left( \frac{1 + S_i(\mu)}{2} \right) \left[ 1 - \sqrt{1 - \left( \frac{4S_i(\mu) \mu^2}{(S_i(\mu) + 1)^2} \right)} \right] = S_{i+1}(\mu)
\]
Thus, using (148), we get that
\[
\min_{\|x\|_2 = 1} \sum_{k=1}^{i+1} \left( y^k x \right)^T \geq \min \{ h_1^*, h_2^* \} \geq \min \{ S_i(\mu) \mu^2, S_{i+1}(\mu) \}
\]
Finally, from Lemma 3.13, we know that \( S_{i+1}(\mu) \leq S_i(\mu) \mu^2 \). Therefore,
\[
\min_{\|x\|_2 = 1} \sum_{k=1}^{i+1} \left( y^k x \right)^T \overset{\|x\|_2}{\geq} S_{i+1}(\mu)
\]
Consequently for any \( x \in Q_{i+1} \), we have
\[
\sum_{k=1}^{i+1} \left( y^k x \right)^T = \left( \sum_{k=1}^{i+1} \left( \left( y^k x \right)^T \right)^2 \right)^{\frac{1}{2}} \overset{\|x\|_2}{\geq} S_{i+1}(\mu) \|x\|_2^2
\]

Now, using Lemmas 3.14 and 3.16, we can show that if \( \{ y^1, \ldots, y^l \} \) satisfy (110) and (111) for each \( i = 1, \ldots, l \), then (115) holds.
Theorem 3.18. Consider the $l$ vectors $\{y^1, \ldots, y^l\} \subset \mathbb{R}^l$ that satisfy (110) and (111) for each $i = 1, \ldots, l$. Then, (115) holds true.

Proof. First of all, since we have assumed that $\{y^1, \ldots, y^l\} \subset \mathbb{R}^l$ satisfy (110) and (111), we know that the requirements of Lemma 3.14 are satisfied and consequently, (124) holds true. Consequently (138) required in Lemma 3.16 holds true for $i = 2$. Therefore, repeatedly applying Lemma 3.16 for $i = 2, \ldots, l - 1$, we finally get that, (139) holds for $i = l - 1$. That is,

$$\min_{x \in \mathbb{R}^l} \frac{\sum_{k=1}^l (y^k^T x)^2}{\|x\|_2^2} \geq S_l(\mu)$$

However, from Lemma 3.12, we know that $\{y^1, \ldots, y^l\}$ are linearly independent vectors in $\mathbb{R}^l$. Therefore, $Q^l = \text{span}\{y^1, \ldots, y^l\} = \mathbb{R}^l$. And hence we get that (115) holds. \qed

Now, having shown our main result regarding the $l$ vectors $\{y^1, \ldots, y^l\} \subset \mathbb{R}^l$ that satisfy (110) and (111), let us turn next to the properties of matrices whose rows satisfy a relaxed form of these conditions.

Theorem 3.19. Consider $m$ row vectors $\{y^1, \ldots, y^m\} \subset \mathbb{R}^l$ of the matrix $Y \in \mathbb{R}^{m \times l}$ defined in (94). As before, let $Q^i := \text{span}\{y^1, \ldots, y^i\}$ for $i = 1, \ldots, l$. Now, given $m \geq l$ and constants $\mu \in (0, 1]$ and $\kappa > 0$, suppose that for each $i = 1, \ldots, l$

$$\|y^i\|_2 \geq \kappa$$

and for each $i = 2, \ldots, l$,

$$\frac{\|\Pi (y^i, (Q^{i-1})^\perp)\|_2^2}{\|y^i\|_2^2} \geq \mu$$

Then, we have

$$\lambda_{\min}(Y^T Y) \geq \kappa^2 S_l(\mu) > 0$$

where $S_l(\mu)$ is the $l$th term of the sequence $\{S_j(\mu)\}_{j \in \mathbb{N}}$ defined in (116).

Proof. We get using (96) and (149) that

$$\lambda_{\min}(Y^T Y) = \min_{x \in \mathbb{R}^l} \frac{\sum_{i=1}^m (y^i^T x)^2}{\|x\|_2^2} \geq \min_{x \in \mathbb{R}^l} \frac{\sum_{i=1}^l (y^i^T x)^2}{\|x\|_2^2} \geq \kappa^2 \left\{ \min_{x \in \mathbb{R}^l} \frac{\sum_{i=1}^l \left( \frac{\|y^i\|_2}{\|y^i\|_2} \right)^2}{\|x\|_2^2} \right\}$$
Let us define
\[ \hat{y}^i = \left( \frac{y^i}{\|y^i\|_2} \right) \text{ for } i = 1, \ldots, l \]

Then it is clear that for each \( i = 1, \ldots, l \),
\[ Q^i = \text{span}\{y^1, \ldots, y^i\} = \text{span}\{\hat{y}^1, \ldots, \hat{y}^i\} \]

Further, it is also easily seen that
\[ \|\hat{y}^i\|_2 = 1 \text{ for } i = 1, \ldots, l \] \hspace{1cm} (153)

Using Property P 3.7 and (150) we get
\[ \|\Pi(\hat{y}^i, Q^{i-1})\|_2 = \left\| \Pi \left( \frac{y^i}{\|y^i\|_2}, Q^{i-1} \right) \right\|_2 = \frac{\|\Pi(y^i, Q^{i-1})\|_2}{\|y^i\|_2} \geq \mu \text{ for } i = 2, \ldots, l \] \hspace{1cm} (154)

Since \( 0 < \mu \leq 1 \), we see from (153) and (154) that the vectors \( \{\hat{y}^1, \ldots, \hat{y}^l\} \subset \mathbb{R}^l \) satisfy the requirements of Theorem 3.18. Therefore, we get that
\[ \min_{x \in \mathbb{R}^l} \sum_{i=1}^l ((\hat{y}^i)^T x)^2 \geq S_l(\mu) \]

where \( S_l(\mu) \) is defined as in (116). Finally, we use this in (152) along with Lemma 3.13, to get that
\[ \lambda^{\min}(Y^T Y) \geq \kappa^2 S_l(\mu) > 0 \] \hspace{1cm} (155)

\[ \square \]

Note: The requirement that the first \( l \) rows of the matrix \( Y \) satisfy (149) and (150) was imposed only for notational convenience. Indeed the bound on the condition number of \( Y^T Y \) can be shown to hold as long as there exists some sequence of \( l \) rows of \( Y \) that satisfy (150). This can be seen from (95) and (96) which show that the maximum and minimum eigenvalues of \( Y^T Y \) are invariant with respect to changes in the ordering of the rows of \( Y \).

Thus, we have established a relationship between the relative positioning of the rows of the matrix \( Y \in \mathbb{R}^{m \times l} \) in \( \mathbb{R}^l \), to the condition number of \( Y^T Y \). The intuition behind the result in Theorem 3.19 is as follows. As we noted immediately after the proof of Lemma 3.11, the minimum eigenvalue of \( Y^T Y \) is small whenever there exists a subspace \( \mathcal{M} \subset \mathbb{R}^l \) of dimension \( l - 1 \) such that all the row vectors \( \{y^1, \ldots, y^m\} \) are close to \( \mathcal{M} \). Consequently, in order to ensure that \( \lambda^{\min}(Y^T Y) \) is bounded away from zero, we must ensure that there exists no subspace \( \mathcal{M} \) such that all the row vectors of \( Y \) are close to \( \mathcal{M} \), i.e., that there exist sufficiently many row vectors of \( Y \) that are “well spread out” in \( \mathbb{R}^l \).
Our notion of the row vectors being well spread-out in \( \mathbb{R}^l \) is given in the conditions (149) and (150). In particular, (149) ensures that there exist \( l \) row vectors \( \{y^1, \ldots, y^l\} \) that lie at least a minimum distance of \( \kappa > 0 \) away from the origin (which clearly belongs to every subspace of \( \mathbb{R}^l \)). Further, note that for each \( i = 2, \ldots, l \), \( (\|\Pi(y^i, Q_i^{-1})\|_2 / \|y^i\|_2) = \sin(\theta_{yi}) \) where \( \theta_{yi} \in [0, \pi/2] \) is the angle between \( y^i \) and \( Q_i^{-1} \). Thus, (150) ensures that for each \( i = 2, \ldots, l \), the angle between the row vector \( y^i \) and the subspace \( Q_i^{-1} \) is at least \( \sin^{-1}(\mu) > 0 \). This in turn ensures that the row vectors \( \{y^1, \ldots, y^l\} \) are well spread out, thus giving us the lower bound in (151).

Indeed there exist other conditions on the rows of \( Y \), that provide similar lower bounds on \( \lambda^\text{min}(Y^T Y) \). The conditions (149) and (150) are particularly appealing because there exists an easy and well known algorithm to verify whether these conditions are satisfied or not. Obviously given \( \{y^1, \ldots, y^m\} \) and \( \kappa > 0 \), it is trivial to check if \( \|y^i\|_2 \geq \kappa \) for \( i = 1, \ldots, l \). Now, consider the result of the Gram-Schmidt orthogonalization process performed on \( \{y^1, \ldots, y^l\} \), particularly the vectors \( \{u^1, \ldots, u^l\} \) defined in (121) and (122). From Property P 3.13, we know that \( u^i = \Pi(y^i, Q_i^{-1}) \) for each \( i = 2, \ldots, l \). Thus, the row vectors \( \{y^1, \ldots, y^l\} \) satisfy (150) if and only if at each step of the Gram-Schmidt process applied to these vectors, we get \( \|u^i\|_2 \geq \mu \|y^i\|_2 \) where \( \mu \in (0, 1] \).

Now, let us apply Theorem 3.19 towards satisfying (101). First, suppose we ensure that for each \( n \in \mathbb{N} \), the design points \( \{x_n + y_n^i : i = 1, \ldots, M_n^l\} \) are picked such that the following assumption is satisfied.

\textbf{A 17.} \hspace{1cm} 1. The number of inner design points \( M_n^l \) is at least \( l \).

2. Given some constant \( \mu_L \in (0, 1] \), the scaled perturbation vectors \( \{\tilde{y}_n^1, \ldots, \tilde{y}_n^l\} \) satisfy for each \( i = 1, \ldots, l \)

\[
\|\tilde{y}_n^i\|_2 \geq \mu_L
\]

(156)

and for each \( i = 2, \ldots, l \)

\[
\frac{\|\Pi(\tilde{y}_n^i, (Q_n^i)^{-1})\|_2}{\|\tilde{y}_n^i\|_2} \geq \mu_L
\]

(157)

where \( Q_n^i := \text{span}\{\tilde{y}_n^1, \ldots, \tilde{y}_n^i\} \) for \( i = 1, \ldots, l \).

Then, using Theorem 3.19, we get that for each \( n \in \mathbb{N} \),

\[
\lambda^\text{min}\left((\tilde{Y}_n^I)^T \tilde{Y}_n^I\right) \geq \mu_L^2 S_l(\mu_L)
\]

(158)

Therefore, (101) is satisfied with \( K_n^I := \mu_L^2 S_l(\mu_L) \). Thus, if for each \( n \in \mathbb{N} \), we pick our inner design points \( \{x_n + y_n^i : i = 1, \ldots, M_n^l\} \) so as to satisfy Assumptions A 16 and A 17 and the corresponding weights \( \{w_n^i : i = 1, \ldots, M_n^l\} \) to satisfy Assumption A 15, then from (97), (104) and (158) we get that for each \( n \in \mathbb{N} \),

\[
\|((\tilde{Y}_n^I)^T W_n^I \tilde{Y}_n^I)^{-1}\|_2 \leq \frac{K_n^I M_n^l}{\mu_L^2 S_l(\mu_L)}
\]

(159)
Similarly, suppose we pick the inner design points \( \{x_n + y_n^i : i = 1, \ldots, M_n^I\} \) for each \( n \in \mathbb{N} \) such that the following assumption is satisfied.

**A 18.**

1. The number of inner design points \( M_n^I \) is at least \( p \).

2. Given some constant \( \mu_Q \in (0, 1] \), the scaled regression vectors \( \{\tilde{z}_n^1, \ldots, \tilde{z}_n^p\} \) satisfy for each \( i = 1, \ldots, p \)

\[
\|\tilde{z}_n^i\|_2 \geq \left(\frac{3}{2}\right) \mu_Q
\]

and for each \( i = 2, \ldots, p \)

\[
\frac{\|\Pi (\tilde{z}_n^i, (R_n^{i-1})^\perp)\|_2}{\|\tilde{z}_n^i\|_2} \geq \mu_Q
\]

where \( R_n^i := \text{span}\{\hat{y}_n^1, \ldots, \hat{y}_n^i\} \) for \( i = 1, \ldots, p \).

Then, using Theorem 3.19, we get that for each \( n \in \mathbb{N} \),

\[
\lambda_{\text{min}} \left( (\tilde{Z}_n^I)^T \tilde{Z}_n^I \right) \geq \left(\frac{3}{2}\right) \mu_Q S_p(\mu_Q)
\]

Therefore, (101) is satisfied with \( K_z^I := \left(\frac{3}{2}\right) \mu_Q S_p(\mu_Q) \). Thus, if for each \( n \in \mathbb{N} \), we pick our inner design points \( \{x_n + y_n^i : i = 1, \ldots, M_n^I\} \) so as to satisfy Assumptions A 16 and 18 and pick the corresponding weights \( \{w_n^i : i = 1, \ldots, M_n^I\} \) to satisfy Assumption A 15, then from (98), (105) and (162) we get that for each \( n \in \mathbb{N} \),

\[
\frac{\| (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I\|_2}{\| ( (\tilde{Z}_n^I)^T W_n^I \tilde{Z}_n^I)^{-1}\|_2} \leq \frac{K_w^I M_n^I}{\mu_Q S_p(\mu_Q)}
\]

Thus, we have two sets of sufficient conditions on the inner design points for each \( n \in \mathbb{N} \), which respectively ensure that (36) in Assumption A 7 and (58) in Assumption A 11 will be satisfied. Next, we develop procedures to pick the design points \( \{x_n + y_n^i : i = 1, \ldots, M_n\} \) to satisfy these conditions. Of course, as we noted earlier, we intend to eventually use the procedures we develop here to construct appropriate regression model functions in trust region algorithms that solve the optimization problem (F). Consequently, in light of Assumption A 1, the procedures we develop to pick \( \{x_n + y_n^i : i = 1, \ldots, M_n\} \) for each \( n \in \mathbb{N} \) so as to satisfy (36) or (58), must at the same time also minimize the number of evaluations of \( F \) required to subsequently evaluate \( \tilde{\nabla}_n f(x_n) \) and \( \tilde{\nabla}_n^2 f(x_n) \).

One obvious solution to reduce the number of evaluations of \( F \) required to construct \( m_n \) for each iteration \( n \) of our trust region algorithms, is to store and reuse sample averages computed in earlier iterations. In general, suppose that we wish to calculate a sample average \( \hat{f}(x, N_2) \) for some \( x \in \mathcal{E} \) and \( N_2 \in \mathbb{N} \). If we have stored the value of \( \hat{f}(x, N_1) \) for some \( 0 < N_1 < N_2 \), then we can calculate \( \hat{f}(x, N_2) \) using

\[
\hat{f}(x, N_2) = \frac{\{N_1 \hat{f}(x, N_1)\} + \sum_{j=N_1}^{N_2} F(x, \tilde{z}_j)}{N_2}
\]
Clearly, such a computation involves only $N_2 - N_1$ extra evaluations of $F$ as against the $N_2$ evaluations required if we do not have $\hat{f}(x, N_1)$ stored. Accordingly, in our trust region algorithms we will maintain a list of at most $M_{\text{max}} < \infty$ sample averages evaluated during the course of the algorithm’s operation and the corresponding points and sample sizes used. In particular, at the beginning of each iteration $n$, for some $\bar{M}_n \leq M_{\text{max}}$ we will have the following (totally) ordered sets.

\[
\begin{align*}
\overline{X}_n & := \{ \overline{x}^k : k = 1, \ldots, \bar{M}_n \} \\
\overline{N}_n & := \{ \overline{N}^k : k = 1, \ldots, \bar{M}_n \} \\
\overline{F}_n & := \{ \hat{f}(\overline{x}^k, \overline{N}^k) : k = 1, \ldots, \bar{M}_n \} \\
\overline{\Phi}_n & := \{ \sigma(\overline{x}^k, \overline{N}^k) : k = 1, \ldots, \bar{M}_n \}
\end{align*}
\]

where for any $x \in E$ and $N \in \mathbb{N}$

\[
\sigma(x, N) := \sqrt{\left( \sum_{j=1}^{N} F(x, \zeta_j)^2 \right) - N \hat{f}(x, N)^2} / N(N - 1)
\]

denotes the standard deviation of the sample average $\hat{f}(x, N)$. Here, for each $k = 1, \ldots, \bar{M}_n$, $\overline{x}^k$ is either a design point $x_n + y_n^k$ or a solution point $x_n$ for an earlier iteration $n^* \in \{1, \ldots, n - 1\}$. If $\overline{x}^k_n = x_n + y_n^k$, for some $n^* \in \{1, \ldots, n - 1\}$, then $\overline{N}_n = N_{n^*}$, $\hat{f}(x_n + y_n^k, N_{n^*})$ and $\sigma(x_n + y_n^k, N_{n^*})$ are stored in $\overline{N}_n$, $\overline{F}_n$ and $\overline{\Phi}_n$ respectively. Similarly, if $\overline{x}^k_n = x_n$, for some $n^* \in \{1, \ldots, n - 1\}$, then $\overline{N}_n = N_0$, $\hat{f}(x_n, N_{n^*})$ and $\sigma(x_n, N_{n^*})$ are stored in $\overline{N}_n$, $\overline{F}_n$ and $\overline{\Phi}_n$ respectively. We store $\sigma(\overline{x}^k_n, \overline{N}^k_n)$ for each $k = 1, \ldots, \bar{M}_n$, for use in a test to adaptively set the sample size $N_{n^*}$ such that (92) may be satisfied.

The advantage of maintaining the lists $\overline{X}_n$, $\overline{N}_n$, $\overline{F}_n$ and $\overline{\Phi}_n$ is clear from (164). Suppose that we are able to set $\overline{x}^k_n$ for some $k \in \{1, \ldots, \bar{M}_n\}$ as the design point $x_n + y_n^k$ for iteration $n$. Then, we can evaluate $\hat{f}(x_n + y_n^k, N_n^i)$ using $\hat{f}(\overline{x}^k_n, \overline{N}^k_n)$ and $N_n^i - \overline{N}^k_n$ extra evaluations of $F$ with (164), whereas we would need $N_n^i$ evaluations of $F$ for a new design point (i.e., a point that is not originally in the set $\overline{X}_n$). Also, note that for any $x \in E$ and $0 < N_1 < N_2$, if we have stored $\sigma(x, N_1)$ and $\hat{f}(x, N_1)$ stored, then after having evaluated $\hat{f}(x, N_2)$ using (164), we can evaluate $\sigma(x, N_2)$ as follows.

\[
\sigma(x, N_2) := \sqrt{N_1(N_1 - 1)\sigma(x, N_1)^2 + \left( \sum_{j=N_1+1}^{N_2} F(x, \tilde{\zeta})^2 \right) - N_2 \hat{f}(x, N_2)^2 - N_1 \hat{f}(x, N_1)^2} / N_2(N_2 - 1)
\]

Consequently, we can also compute $\sigma(x_n + y_n^i, N_n^i)$ using $\sigma(\overline{x}^k_n, \overline{N}^k_n)$, $\hat{f}(\overline{x}^k_n, \overline{N}^k_n)$, $\hat{f}(x_n + y_n^i, N_n^i)$ and $\overline{N}^k_n - N_n^i$ evaluations of $F$.

Thus, in each iteration $n$ of our trust region algorithms, we will pick as many design points as possible from $\overline{X}_n$ and evaluate sample averages the fewest possible number of “new” (i.e., not in $\overline{X}_n$) design points required to satisfy either Assumption A17 or Assumption A18 depending on whether we wish to satisfy (36) or (58). Also, whenever we evaluate a sample average at a point not in $\overline{X}_n$, we will insert this “new” point
into $\mathcal{X}_n$ and the corresponding sample size, sample average and standard deviation into $\mathcal{N}_n$, $\mathcal{F}_n$ and $\mathcal{\Phi}_n$ respectively. Finally, at the end of iteration $n$, we will pick at most $M_{\text{max}}$ appropriate points from $\mathcal{X}_n$ (for example, the points in $\mathcal{X}_n$ that lie closest to $x_{n+1}$ in the Euclidean norm) and define the set of these points to be $\mathcal{X}_{n+1}$ and define the corresponding sets of sample sizes, sample averages and standard deviations to be $\mathcal{N}_{n+1}$, $\mathcal{F}_{n+1}$ and $\mathcal{\Phi}_{n+1}$ respectively. Thus, $\mathcal{X}_n$, $\mathcal{N}_n$, $\mathcal{F}_n$ and $\mathcal{\Phi}_n$ start as empty sets for $n = 1$. As our algorithms progress, these sets grow to contain $M_{\text{max}}$ elements each and $M_n$ remains equal to $M_{\text{max}}$ thereafter.

Accordingly, in this section, we will assume that for each $n \in \mathbb{N}$, the ordered sets $\mathcal{X}_n$, $\mathcal{N}_n$, $\mathcal{F}_n$ and $\mathcal{\Phi}_n$ as defined in (165) through (168) respectively, are available to us with $\mathcal{M}_n \leq M_{\text{max}}$. We will provide algorithms to pick as many inner design points as possible from $\mathcal{X}_n$ required to satisfy Assumptions A 17 or A 18. Further, we provide algorithms to appropriately pick the required number of “new” inner design points, if sufficiently many points do not already exist in $\mathcal{X}_n$ to satisfy Assumptions A 17 or A 18. Using these algorithms, we then provide methods to pick the inner and outer design points for each $n \in \mathbb{N}$, so as to satisfy Assumptions A 16 and either Assumption A 17 or Assumption A 18. Since the algorithms we discuss in this section involve the manipulation of sets such as $\mathcal{X}_n$, we briefly describe the notation required for such manipulations first.

1. For our purposes, we define an ordered set $(\mathcal{A}, f_{\mathcal{A}})$ as an ordered pair consisting of a set $\mathcal{A}$ of $m \in \mathbb{Z}$ unique objects (for some $m \geq 0$) along with a bijective mapping $f_{\mathcal{A}}: \{1, \ldots, m\} \longrightarrow \mathcal{A}$. If the set $\mathcal{A} = \emptyset$ i.e. $m = 0$, then we define the ordered set $(\mathcal{A}, f_{\mathcal{A}})$ also to be empty. Thus, an ordered set for us, is essentially a finite set on which a total order has been imposed. For notational brevity however, we will not explicitly refer to the bijection $f_{\mathcal{A}}$ while referring to an ordered set $(\mathcal{A}, f_{\mathcal{A}})$ since in most cases, the ordering of the elements will be clear from context. Instead whenever we refer to a set $\mathcal{A} := \{a, b, c, \ldots\}$, it is understood that the the elements of the set $\mathcal{A}$ have been written in the order defined by $f_{\mathcal{A}}$, i.e., that $a = f_{\mathcal{A}}(1)$, $b = f_{\mathcal{A}}(2)$ and so on. Further, so as not to explicitly involve $f_{\mathcal{A}}$ in the notation, for each $i \in \{1, \ldots, m\}$, we will denote $f_{\mathcal{A}}(i)$ by $\mathcal{A}(i)$ and refer to it as the $i^{\text{th}}$ element of the set $\mathcal{A}$. Finally, we let $|\mathcal{A}|$ denote the number of elements in the set $\mathcal{A}$. If $|\mathcal{A}| = 0$, then we say that the ordered set $\mathcal{A}$ is empty and write $\mathcal{A} = \emptyset$.

2. In contexts where the ordering of the elements in a set $\mathcal{A}$ is not important, we will treat $\mathcal{A}$ as a regular set and perform common set operations on it.

3. Suppose that $\mathcal{A} = \{a_1, \ldots, a_m\}$ is an ordered set of $m$ elements for some $m > 0$. Let $\mathcal{K} = \{i_1, \ldots, i_j\}$ be an ordered set of positive integers containing $j \leq m$ elements such that $\mathcal{K}(i) \in \{1, \ldots, m\}$ for each $k = 1, \ldots, j$. Then we say that $\mathcal{K}$ is an index set for $\mathcal{A}$ and define $\mathcal{A}(\mathcal{K})$ to be the ordered set $\{a_{i_1}, a_{i_2}, \ldots, a_{i_j}\}$.

4. We define next, some operations that can be performed on an ordered set.
We can set the object at position \( j \) (for \( 1 \leq j \leq m \)) in \( \mathcal{A} = \{a_1, \ldots, a_m\} \) to be \( b \). After such an operation is performed, \( \mathcal{A} \) is defined as the set of \( m + 1 \) elements, \( \{a_1, \ldots, a_{j-1}, b, a_{j+1}, \ldots, a_m\} \).

We can insert the object \( b \) at position \( j \) for some \( 1 \leq j \leq m + 1 \) into the set \( \mathcal{A} \). After the insertion operation, \( \mathcal{A} \) is a set of \( m + 1 \) objects defined as follows.

\[
\mathcal{A} := \begin{cases} 
\{b, a_1, a_2, \ldots, a_m\} & \text{if } j = 1 \\
\{a_1, \ldots, a_{j-1}, b, a_j, \ldots, a_m\} & \text{if } 2 \leq j \leq m \\
\{a_1, \ldots, a_m, b\} & \text{if } j = m + 1
\end{cases}
\]

We can delete the \( j^{th} \) object from the set \( \mathcal{A} \) for some \( 1 \leq j \leq m \). After the deletion operation \( \mathcal{A} \) is a set of \( m - 1 \) elements defined as follows.

\[
\mathcal{A} := \begin{cases} 
\{a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_m\} & \text{if } m > 1 \\
\emptyset & \text{if } m = 1
\end{cases}
\]

In keeping with the terminology developed earlier in Section 3, for each \( k = 1, \ldots, M_n \), we will refer to the vectors \((\bar{X}_n(k) - x_n)\) and \((\bar{X}_n(k) - x_n)/\delta_n\) as the perturbation and regression vectors respectively, corresponding to \( \bar{X}_n(k) \). Similarly, we will refer to \(( (\bar{X}_n(k) - x_n)/\delta_n ) \) and \(( (\bar{X}_n(k) - x_n)/\delta_n^2 ) \) as the scaled perturbation and regression vectors respectively, corresponding to \( \bar{X}_n(k) \). Now, we start by providing an algorithm that separates the points in \( \bar{X}_n \) lying within and outside the design region \( D_n \) and also identifies as many points from \( \bar{X}_n \) as possible that can be used as design points that satisfy Assumption A 17.

**Algorithm 2.** Let the ordered set \( \bar{X}_n \) containing \( M_n \) (where \( M_n \leq M_{\text{max}} \)) entries be given. Also, we assume that the design region radius \( \delta_n > 0 \), the current solution \( x_n \) and a constant \( \mu_L \in (0,1] \) are known. Initialize the index sets \( \bar{K}_n^g = \bar{K}_n^b = \bar{K}_n = \emptyset \) and the corresponding counters \( M_n^g = M_n^b = M_n^o = 0 \). Also initialize the sets \( \hat{\mathcal{Y}}_n = \hat{\mathcal{Z}}_n = \mathcal{U}_n = \emptyset \). Repeat the following steps for each \( k \in \{1, \ldots, M_n\} \).

**Step 1:** Set \( \hat{y}_n \leftarrow (\bar{X}_n(k) - x_n)/\delta_n \) and compute \((\hat{y}_n)^Q\) (as defined in (16)).

**Step 2:** Insert \( \hat{y}_n \) into the set \( \mathcal{Y}_n \) and \((\hat{y}_n)^Q\) into the set \( \mathcal{Z}_n \) both at position \( k \).

**Step 3:** If \( 0 < \|\hat{y}_n\|_2 \leq 1 \) then
- If \( 0 \leq \bar{M}_n^g < l \), then
  - Set
  \[
  u_n \leftarrow \begin{cases} 
  \hat{y}_n & \text{if } \bar{M}_n^g = 0 \\
  \hat{y}_n - \sum_{j=1}^{\bar{M}_n^g} (\hat{y}_n^T \mathcal{U}_n(j)) \mathcal{U}_n(j) & \text{otherwise}
  \end{cases}
  \]

(171)
- If \( \| \tilde{y}_n \|_2 < \mu_L \) or \( \| u_n \|_2 < \mu_L \| \tilde{y}_n \|_2 \), then set \( M^b_n \leftarrow M^b_n + 1 \) and insert \( k \) into the set \( \mathcal{K}^b_n \) at the position \( M^b_n \).

- If \( \| \tilde{y}_n \|_2 \geq \mu_L \) and \( \| u_n \|_2 \geq \mu_L \| \tilde{y}_n \|_2 \) then
  
  * Set \( M^G_n \leftarrow M^G_n + 1 \) and insert \( k \) into the set \( \mathcal{K}^G_n \) at the position \( M^G_n \).

  * Insert \( (u_n/\| u_n \|_2) \) into the set \( \mathcal{U}_n \) at the position \( M^G_n \).

- If \( M^G_n = l \), then set \( M^G_n \leftarrow M^G_n + 1 \) and insert \( k \) into the set \( \mathcal{K}^G_n \) at the position \( M^G_n \).

**Step 4:** If \( \| \tilde{y}_n \|_2 > 1 \), then \( \mathcal{M}^G_n \leftarrow \mathcal{M}^G_n + 1 \) and insert \( k \) into the set \( \mathcal{K}^G_n \) at the position \( M^G_n \).

It is easily seen from the operation of Algorithm 2 that the various sets generated in the algorithm satisfy the following properties.

**P 3.14.** For each \( k = 1, \ldots, \mathcal{M}_n \), \( \hat{y}_n(k) \) and \( \hat{Z}_n(k) \) are respectively, the scaled perturbation and regression vectors for the point \( \mathcal{X}_n(k) \), evaluated with respect to \( x_n \).

**P 3.15.** 1. The sets \( \mathcal{K}^G_n, \mathcal{K}^b_n, \text{ and } \mathcal{K}^G_n \), containing \( \mathcal{M}^G_n, \mathcal{M}^b_n, \text{ and } \mathcal{M}^G_n \) elements respectively, satisfy \( \mathcal{K}^G_n \cap \mathcal{K}^b_n = \mathcal{K}^b_n \cap \mathcal{K}^G_n = \mathcal{K}^G_n \cap \mathcal{K}^G_n = \{1, \ldots, \mathcal{M}_n \} \). Consequently, we have \( \mathcal{M}^G_n + \mathcal{M}^b_n + \mathcal{M}^G_n = \mathcal{M}_n \).

2. \( \| \hat{y}_n(\mathcal{K}^G_n(i)) \|_2 \leq 1 \) for each \( i = 1, \ldots, \mathcal{M}^G_n \) and \( \| \hat{y}_n(\mathcal{K}^G_n(i)) \|_2 \leq 1 \) for each \( i = 1, \ldots, \mathcal{M}^b_n \). Consequently, we get that \( \mathcal{X}_n(\mathcal{K}^G_n) \cup \mathcal{X}_n(\mathcal{K}^b_n) \subset \mathcal{D}_n \), where \( \mathcal{D}_n \) is the design region for iteration \( n \) defined as in (26).

Thus, the points in \( \mathcal{X}_n(\mathcal{K}^G_n) \) and \( \mathcal{X}_n(\mathcal{K}^b_n) \) together form a set of potential inner design points for iteration \( n \). Similarly \( \| \hat{y}_n(\mathcal{K}^G_n(i)) \|_2 \geq 1 \) for each \( i = 1, \ldots, \mathcal{M}^b_n \) and hence the points in \( \mathcal{X}_n(\mathcal{K}^G_n) \in \mathbb{R}^l \setminus \mathcal{D}_n \) form a set of potential outer design points for iteration \( n \).

**P 3.16.** 1. If \( \mathcal{M}^G_n \geq 1 \), the vectors in the set \( \mathcal{U}_n \) are the Gram-Schmidt vectors corresponding to those in \( \hat{y}_n(\mathcal{K}^G_n) \). That is, if \( \mathcal{M}^G_n \geq 1 \), then

\[
\mathcal{U}_n(1) = \frac{\hat{y}_n(\mathcal{K}^G_n(1))}{\| \hat{y}_n(\mathcal{K}^G_n(1)) \|_2} \tag{172}
\]

and if \( \mathcal{M}^G_n \geq 2 \) then for \( i = 2, \ldots, \mathcal{M}^G_n \),

\[
\mathcal{U}_n(i) = \frac{\hat{y}_n(\mathcal{K}^G_n(i)) - \sum_{j=1}^{i-1} \left( \hat{y}_n(\mathcal{K}^G_n(i))^T \mathcal{U}_n(j) \right) \mathcal{U}_n(j)}{\| \hat{y}_n(\mathcal{K}^G_n(i)) - \sum_{j=1}^{i-1} \left( \hat{y}_n(\mathcal{K}^G_n(i))^T \mathcal{U}_n(j) \right) \mathcal{U}_n(j) \|_2} \tag{173}
\]

2. If \( \mathcal{M}^G_n \geq 1 \), then for each \( i = 1, \ldots, \mathcal{M}^G_n \),

\[
\| \hat{y}_n(\mathcal{K}^G_n(i)) \|_2 \geq \mu_L \tag{174}
\]
Further, if $|\overline{M}_n^G| \geq 2$ then we get from Property P 3.13 that for $i = 2, \ldots, |\overline{M}_n^G|,$

$$
\| \Pi \left( \tilde{Y}_n(\mathcal{K}_n^G(i)), \left( \text{span} \left\{ \tilde{Y}_n(\mathcal{K}_n^G(1)), \ldots, \tilde{Y}_n(\mathcal{K}_n^G(i-1)) \right\} \right)^\perp \right) \|_2
\leq \| \tilde{Y}_n(\mathcal{K}_n^G(i)) \|_2
= \| \tilde{Y}_n(\mathcal{K}_n^G(i)) - \sum_{j=1}^{i-1} \left( \tilde{Y}_n(\mathcal{K}_n^G(i))U_n(j) \right) U_n(j) \|_2
\geq \mu_L \quad (175)
$$

Indeed, it is easily seen that Algorithm 2 is only a slightly modified form of the Gram-Schmidt orthogonalization process; the only difference being that in Algorithm 2, the set of vectors on which the Gram-Schmidt process is performed i.e., $\mathcal{Y}_n(\mathcal{K}_n^G)$, is chosen from $\mathcal{Y}_n$ as the algorithm progresses, so as to ensure that the vectors in $\mathcal{Y}_n(\mathcal{K}_n^G)$ satisfy (174) and (175). Consequently, from Property P 3.16, it is clear that if $|\overline{M}_n^G| \geq 1$ and we set $x_n + y_n^G = \mathcal{X}_n(\mathcal{K}_n^G(i))$ for each $i = 1, \ldots, |\overline{M}_n^G|$, then (156) is satisfied for $i = 1, \ldots, |\overline{M}_n^G|$ and if $|\overline{M}_n^G| \geq 2$, then (157) is satisfied for $i = 2, \ldots, |\overline{M}_n^G|$.

However, recall that $l$ inner design points satisfying (156) and (157) are required to satisfy Assumption A 17 and it is clearly possible that after Algorithm 2 is executed we get fewer than $l$ points in $\mathcal{X}_n(\mathcal{K}_n^G)$, i.e., $|\overline{M}_n^G| < l$. Thus, we describe a procedure next, that inserts an appropriate point into $\mathcal{X}_n$, such that $|\overline{M}_n^G|$ is incremented by one and the Properties P 3.14 through P 3.16 continue to hold.

**Algorithm 3.** Let us assume that we have knowledge of the design region radius $\delta_n$, the sets $\mathcal{X}_n$, $\mathcal{Y}_n$, $\mathcal{Z}_n$, $\mathcal{N}_n$, $\mathcal{F}_n$, $\mathcal{F}_n$ and $\mathcal{U}_n$, the index sets $\mathcal{K}_n^G$, $\mathcal{K}_n^B$ and $\mathcal{K}_n^O$ and a constant $\mu_L \in (0,1)$, such that Properties P 3.14 through P 3.16 are satisfied. Further, we assume that $|\overline{M}_n^G| < l$ and let $\{e_1, \ldots, e_l\} \subset \mathbb{R}^l$ represent the standard orthonormal basis for $\mathbb{R}^l$.

**Step 1:** If $|\overline{M}_n^G| = 0$, then set $\tilde{y}_n \leftarrow e_1$.

**Step 2:** If $|\overline{M}_n^G| \geq 1$, then

- Find $i^* \in \{1, \ldots, l\}$ such that

$$
\sum_{j=1}^{i^*} (e_i^T U_n(j))U_n(j) = u_n = e_i^* - \sum_{j=1}^{i^*} (e_i^T U_n(j))U_n(j) \quad \text{and} \quad \|u_n\|_2 > 0 \quad (176)
$$

- Set $\tilde{y}_n \leftarrow (u_n / \|u_n\|_2)$.

**Step 3:** Set $\overline{M}_n \leftarrow |\overline{M}_n| + 1$.

**Step 4:** Insert $\tilde{y}_n$ into $\mathcal{Y}_n$, $(\tilde{y}_n, (\tilde{y}_n)^T)$ into $\mathcal{Z}_n$ and $x_n + \delta_n \tilde{y}_n$ into $\mathcal{X}_n$ all at position $|\overline{M}_n|$.

**Step 5:** Insert the value 0 into $\mathcal{F}_n$, $\mathcal{F}_n$ and $\mathcal{N}_n$ all at position $|\overline{M}_n|$.
Step 6: Set $\mathcal{M}_n^G \leftarrow \mathcal{M}_n^G + 1$ and insert $\mathcal{M}_n^G$ into the set $\mathcal{K}_n^G$ at position $\mathcal{M}_n^G$.

Step 7: Insert $\hat{y}_n$ into the set $\mathcal{U}_n$ at the position $\mathcal{M}_n^G$.

Algorithm 3 essentially computes a unit vector $\hat{y}_n \in (\text{span}(\hat{y}_n(\mathcal{K}_n^G)))^\perp$ sets $x_n + \delta_n \hat{y}_n$ as the new design point to be added to $\mathcal{X}_n$. Note that the sample size, sample average and standard deviation of the new point are each set to be zero. We will provide methods to set these quantities to their proper values in Section 3.2.2. The following lemma shows that as long as $\mathcal{M}_n^G < l$ before the execution of Algorithm 3, the algorithm will successfully add a point to $\mathcal{X}_n$ such that the updated set of scaled perturbations $\hat{y}_n(\mathcal{K}_n^G)$, continues to satisfy (174) and (175)

**Lemma 3.20.** Suppose that the sets $\mathcal{X}_n, \hat{y}_n, \hat{Z}_n, \mathcal{N}_n, \mathcal{F}_n, \Phi_n$ and $\mathcal{U}_n$ and the index sets $\mathcal{K}_n^G$, $\mathcal{K}_n^B$ and $\mathcal{K}_n^I$ satisfy Properties P 3.14 through P 3.16 and we have $\mathcal{M}_n^G < l$ before the execution of Algorithm 3. Then, the algorithm is well defined and upon its termination, $\mathcal{M}_n^G$ is incremented by one and Properties P 3.14 through P 3.16 continue to hold.

**Proof.** In order to show that Algorithm 2 is well defined, we need only to show that Steps 2 in the algorithm well defined since it is trivial to see that the rest of the steps are well defined. In particular, for Step 2 to be well defined, we need to show that if $\mathcal{M}_n^G < l$, there exists some $i^* \in \{1, \ldots, l\}$ such that (176) holds. Suppose $1 \leq \mathcal{M}_n^G < l$ before running Algorithm 3. Then we know from Property P 3.16 that $\mathcal{U}_n$ contains the Gram-Schmidt orthonormal vectors corresponding to those in $\hat{y}_n(\mathcal{K}_n^G)$ where $\mathcal{M}_n^G < l$. Therefore,

$$\dim(\text{span}\{\mathcal{U}_n\}) = \dim(\text{span}\{\hat{y}_n(\mathcal{K}_n^G)\}) = \mathcal{M}_n^G < l$$

Since $\dim(\text{span}\{e^1, \ldots, e^l\}) = l$, clearly there exists some $i^* \in \{1, \ldots, l\}$ such that $e^{i^*} \notin \text{span}\{\mathcal{U}_n\}$, i.e., such that

$$\|u_n\|_2 > 0, \quad \text{where} \quad u_n = e^{i^*} - \sum_{j=1}^{\mathcal{M}_n^G} (e^{i^*})^T \mathcal{U}_n(j) \mathcal{U}_n(j)$$

Therefore Algorithm 3 is able to set $\hat{y}_n = (u_n/\|u_n\|_2)$ and hence Step 2 of Algorithm 3 is well defined.

Next we show that if Properties P 3.14, and P 3.15 were satisfied before the execution of Algorithm 3, they continue to be satisfied after its execution. It is easily seen that upon termination of Algorithm 3, $\mathcal{M}_n$ is incremented by one and $\hat{y}_n(\mathcal{M}_n) = \hat{y}_n$, $\hat{Z}_n(\mathcal{M}_n) = \begin{pmatrix} \hat{y}_n \\ (\hat{y}_n)^Q \end{pmatrix}$ and $\mathcal{X}_n(\mathcal{M}_n) = x_n + \delta_n \hat{y}_n$. Since no other changes are made to $\mathcal{X}_n$, $\hat{y}_n$ and $\hat{Z}_n$, and Property P 3.14 held before the execution of the algorithm, it is clear that this property holds after its execution.

Similarly, upon termination of Algorithm 3, $\mathcal{M}_n^G$ is incremented by one and we get $\mathcal{K}_n^G(\mathcal{M}_n^G) = \mathcal{M}_n^G$. Further, note that since $\mathcal{M}_n$ was incremented by one, $\mathcal{M}_n^G \notin \mathcal{K}_n^B \cup \mathcal{K}_n^G$. That is, a new entry($\mathcal{M}_n$) is inserted only into $\mathcal{K}_n^G$ and no changes are made to $\mathcal{K}_n^B$ or $\mathcal{K}_n^I$. Consequently, since the first part of Property P 3.15
was assumed to hold before the execution of Algorithm 3, it continues to hold after the execution of the algorithm. Further, since \( \|\hat{\gamma}_n(\mathcal{M}_n)\|_2 = \|\hat{\gamma}_n\|_2 = 1 \) (where \( \hat{\gamma}_n \) is calculated in either Step 1 or Step 2), it is clear that \( \mathcal{X}_n(\mathcal{M}_n) \in \mathcal{P}_n \). Therefore, after the execution of the algorithm \( \mathcal{X}_n(\mathcal{K}_n^G) \subset \mathcal{P}_n, \mathcal{X}_n(\mathcal{K}_n^B) \subset \mathcal{P}_n \) and \( \mathcal{X}_n(\mathcal{K}_n^G) \subset \mathbb{R}^l \setminus \mathcal{P}_n \) and the second part of Property P 3.15 continues to hold.

Next, let us show that if Property P 3.16 held before the execution of Algorithm 3, then it holds after its execution. Note again, that upon termination of the algorithm, the value of \( \mathcal{M}_n^G \) is incremented by one and \( \mathcal{U}_n(\mathcal{M}_n^G) = \hat{\gamma}_n(\mathcal{K}_n^G(\mathcal{M}_n^G)) = \hat{\gamma}_n \). First, if \( \mathcal{M}_n^G = 0 \) before the execution of Algorithm 3, then after its execution \( \mathcal{M}_n^G = 1 \) and \( \mathcal{U}_n(\mathcal{M}_n^G) = \hat{\gamma}_n(\mathcal{K}_n^G(\mathcal{M}_n^G)) = e^1 \). Thus, it is clear that (172) holds. Further, since \( \|\hat{\gamma}_n(\mathcal{M}_n^G)\|_2 = 1 \), it is easily seen that (174) is satisfied for \( i = 1 \).

Finally, suppose that \( \mathcal{M}_n^G \geq 1 \) before the execution of Algorithm 3. Then, \( \mathcal{M}_n^G \) is incremented by one and consequently, \( \mathcal{M}_n^G \geq 2 \) after its execution. Since we assumed that the first part of Property P 3.16 held before the algorithm was executed, we know that after it is executed, (172) holds and if \( \mathcal{M}_n^G \geq 3 \) after the execution of the algorithm then (173) holds for \( i = 2, \ldots, \mathcal{M}_n^G - 1 \). Therefore, from Property P 3.11, we get that \( \{\mathcal{U}_n(1), \ldots, \mathcal{U}_n(\mathcal{M}_n^G - 1)\} \) is an orthonormal set. Using this in (176), we get that

\[
\begin{aligned}
  u_i^T \mathcal{U}_n(j) &= (e^i)^T \mathcal{U}_n(j) - (e^i)^T \mathcal{U}_n(j) = 0 \quad \text{for each} \quad j = 1, \ldots, \mathcal{M}_n^G - 1
\end{aligned}
\]

Thus, since \( \hat{\gamma}_n = (u_n/\|u_n\|_2) \), we get that \( \|\hat{\gamma}_n\|_2 = 1 \) and

\[
\begin{aligned}
  (\hat{\gamma}_n)^T \mathcal{U}_n(j) &= 0 \quad \text{for each} \quad j = 1, \ldots, \mathcal{M}_n^G - 1
\end{aligned}
\]

(177)

Since \( \mathcal{U}_n(\mathcal{M}_n^G) = \hat{\gamma}_n(\mathcal{K}_n^G(\mathcal{M}_n^G)) = \hat{\gamma}_n \), we get using (177) and \( \|\hat{\gamma}_n\|_2 = 1 \), that

\[
\begin{aligned}
  \mathcal{U}_n(\mathcal{M}_n^G) = \hat{\gamma}_n &= \frac{\hat{\gamma}_n - \sum_{j=1}^{\mathcal{M}_n^G-1} (\hat{\gamma}_n^T \mathcal{U}_n(j)) \mathcal{U}_n(j)}{\|\hat{\gamma}_n - \sum_{j=1}^{\mathcal{M}_n^G-1} (\hat{\gamma}_n^T \mathcal{U}_n(j)) \mathcal{U}_n(j)\|_2} \\
  &= \frac{\hat{\gamma}_n(\mathcal{K}_n^G(\mathcal{M}_n^G)) - \sum_{j=1}^{\mathcal{M}_n^G-1} (\hat{\gamma}_n(\mathcal{K}_n^G(\mathcal{M}_n^G))^T \mathcal{U}_n(j)) \mathcal{U}_n(j)}{\|\hat{\gamma}_n(\mathcal{K}_n^G(\mathcal{M}_n^G)) - \sum_{j=1}^{\mathcal{M}_n^G-1} (\hat{\gamma}_n(\mathcal{K}_n^G(\mathcal{M}_n^G))^T \mathcal{U}_n(j)) \mathcal{U}_n(j)\|_2}
\end{aligned}
\]

Therefore, (173) holds for \( i = \mathcal{M}_n^G \). Thus, the first part of Property P 3.16 holds after the execution of Algorithm 3.

Similarly, since the second part of Property P 3.16 was assumed to hold before the execution of Algorithm 3, we get that after its execution (174) holds for \( i = 1, \ldots, \mathcal{M}_n^G - 1 \) and if \( \mathcal{M}_n^G \geq 3 \), (175) holds for \( i = 2, \ldots, \mathcal{M}_n^G - 1 \). Thus, in order to show that the second part of Property P 3.16 holds after the execution of Algorithm 3, we only have to show that (174) and (175) hold for \( i = \mathcal{M}_n^G \). Obviously, since

\[
\|\hat{\gamma}_n(\mathcal{K}_n^G(\mathcal{M}_n^G))\|_2 = \|\hat{\gamma}_n\|_2 = 1 \geq \mu_L, \quad (174) \text{ holds for } i = \mathcal{M}_n^G \]

Further, we get using Property P 3.13 and
Therefore (175) holds for \( i = \overline{M}_n^G \) and consequently, the second part of Property P 3.16 continues to hold upon termination of Algorithm 3.

Now if we have the set \( \mathcal{X}_n \) (and the corresponding \( \mathcal{N}_n, \mathcal{F}_n \) and \( \Phi_n \)) for each \( n \in \mathbb{N} \), we can use Algorithms 2 and 3 to find a set of design points \( \{x_n + y_n^i : i = 1, \ldots, M_n\} \) such that Assumptions A 16 and A 17 are satisfied.

**Algorithm 4.** Let us assume that for each \( n \in \mathbb{N} \), we have the sets \( \mathcal{X}_n, \mathcal{N}_n, \mathcal{F}_n \) and \( \Phi_n \) with \( \overline{M}_n \leq M_{\text{max}} \).

Initialize the index sets \( K_n^G = K_n^A = K_n^O = \emptyset \) and the corresponding counters \( M_n^G = M_n^A = M_n^O = 0 \). Let the constant \( l \geq M_{\text{max}}^l < \infty \) be given.

**Step 1**: Run Algorithm 2 once. Then, run Algorithm 3 as many times as necessary to obtain the sets \( \hat{\mathcal{Y}}_n, \hat{\mathcal{Z}}_n, \mathcal{U}_n \) and the index sets \( \overline{K}_n^G, \overline{K}_n^A, \overline{K}_n^O \) with \( \overline{M}_n = l, \overline{M}_n^A \) and \( \overline{M}_n^O \) elements respectively. Note that Algorithm 3 has to be run at most \( l \) times since the value of \( \overline{M}_n^G \) is incremented by one whenever Algorithm 3 is run.

**Step 2**: Set \( \mathcal{K}_n^G \leftarrow \overline{K}_n^G \) and \( M_n^G \leftarrow \overline{M}_n^G \).

**Step 3**: Set the index set \( K_n^A \) with \( 0 \leq M_n^A \leq \max \{ \overline{M}_n^A, M_{\text{max}}^l - l \} \) elements, such that it contains a subset of the elements in \( K_n^A \) in some order (the order can be chosen arbitrarily). For example, we may set \( K_n^A \leftarrow \emptyset \) or \( K_n^A \leftarrow \overline{K}_n^A \) or \( K_n^A \leftarrow \{ i \in \overline{K}_n^A : \mathcal{N}_n(\overline{K}_n^A(i)) \geq \alpha N_n^0 \} \) for some constant \( \alpha \in (0, 1] \).

**Step 4**: Set the index set \( K_n^O \) with \( 0 \leq M_n^O \leq \overline{M}_n^G \) elements, such that it contains a subset of the elements in \( K_n^O \) in some order (the order can be chosen arbitrarily). For example, we may set \( K_n^O \leftarrow \emptyset \) or \( K_n^O \leftarrow \overline{K}_n^O \) or \( K_n^O \leftarrow \{ i \in \overline{K}_n^O : \| \hat{\mathcal{Y}}_n(\overline{K}_n^O(i)) \|_2 \leq \alpha \delta \} \) for some constant \( \alpha \delta > 1 \).

**Step 5**: For each \( i = 1, \ldots, M_n^G \), set \( x_n + y_n^i \leftarrow \overline{X}_n(\mathcal{K}_n^G(i)) \).

**Step 6**: If \( M_n^G > 0 \), then

- For each \( i = M_n^G + 1, \ldots, M_n^G + M_n^A \), set \( x_n + y_n^i \leftarrow \overline{X}_n(\mathcal{K}_n^G(i - M_n^G)) \).

**Step 7**: If \( M_n^O > 0 \), then

\[
\left\| \hat{\mathcal{Y}}_n(\mathcal{K}_n^G(\mathcal{M}_n^G)), \left( \text{span} \{ \hat{\mathcal{Y}}_n(\mathcal{K}_n^G(1)), \ldots, \hat{\mathcal{Y}}_n(\mathcal{K}_n^G(\mathcal{M}_n^G - 1)) \} \right) \right\|_2 = \frac{\| \hat{y}_n - \sum_{j=1}^{\overline{M}_n^G} \left( (\hat{y}_n)^T \mathcal{U}_n(j) \right) \mathcal{U}_n(j) \|_2}{\| \hat{y}_n \|_2} = 1 \geq \mu_L
\]
For each \(i = M_n^G + M_n^B + 1, \ldots, M_n^G + M_n^B + M_n^O\), set \(x_n + y_n^i \leftarrow \mathbf{\nabla}_n (K_n^G (i - M_n^G - M_n^B))\).

If Algorithm 4 is run for each \(n \in \mathbb{N}\), then from Lemma 3.20, it is easily seen that the following statements hold.

1. The number of entries in \(K_n^G\) is equal to \(l\) and the number of entries in \(K_n^B\) is at most \(M_{\text{max}}^I - l\). Thus, since \(\mathbf{\nabla}_n (K_n^G) \cup \mathbf{\nabla}_n (K_n^B)\) form the set of inner design points, it is clear that there exist at most \(M_{\text{max}}^I\) inner design points for each \(n \in \mathbb{N}\). Therefore, Assumption A 16 is satisfied for each \(n \in \mathbb{N}\).

2. From Lemma 3.20, we know that Properties P 3.14 through P 3.16 are satisfied after the execution of Step 1 in Algorithm 4. Thus, from Property P 3.16 we know that the \(l\) vectors in \(\hat{\mathbf{y}}_n (K_n^G)\) satisfy (174) and (175) and from Property P 3.14, we know that these vectors are the scaled perturbation vectors corresponding to the points in \(\mathbf{\nabla}_n (K_n^G)\).

Finally, we note from Steps 2 and 5 of Algorithm 4 that the points in \(\mathbf{\nabla}_n (K_n^G) = \mathbf{\nabla}_n (K_n^G)\) are chosen to be the first \(l\) design points. Therefore, we get that (156) and (157) is satisfied by the first \(l\) design points and consequently Assumption A 17 is satisfied.

Thus, for each \(n \in \mathbb{N}\), as long as the design points are chosen using Algorithm 4 and the weights \(\{w_n^i : i = 1, \ldots, M_n^I\}\) are chosen so as to satisfy Assumption A 15, we get from (97), (104) and (158) that (159) holds, i.e. (36) in Assumption A 7 holds with \(K_\lambda^I = (K_n^I M_{\text{max}}^I / \mu^L S_{\theta}(\mu_L))\) for each \(n \in \mathbb{N}\).

Next, we look at methods to pick \(\{x_n + y_n^i : i = 1, \ldots, M_n^I\}\) so as to satisfy Assumption A 18. The algorithms we will devise to find \(p\) design points satisfying the conditions (160) and (161), will be similar to Algorithms 2 and 3. That is, we will first provide an algorithm to identify as many points as possible from the set \(\mathbf{\nabla}_n\) that can be used to satisfy (160) and (161) and then provide another procedure find the required number of new design points that can be inserted into \(\mathbf{\nabla}_n\) so as to ensure that these requirements are satisfied.

To this end, we start with a Lemma that shows that the \(l\) scaled regression vectors in \(\hat{Z}_n (K_n^G)\), obtained after running Algorithms 2 once and 3 at most \(l\) times, can be set as the first \(l\) design points satisfying (160) and (161).

**Lemma 3.21.** Consider the \(l\) scaled regression vectors in the set \(\hat{Z}_n (K_n^G)\) obtained after the execution of Algorithm 2 once and Algorithm 3 as many times as required to get \(\mathbf{\nabla}_n^G = l\). We have for each \(i = 1, \ldots, l\),

\[
\left\| \hat{Z}_n (K_n^G (i)) \right\|_2 \geq \mu_L = \left( \frac{\sqrt{3}}{2} \right) \left( \frac{2}{\sqrt{3}} \right) \mu_L
\]
and for $i = 2, \ldots, l$,
\[
\left\| \Pi \left( \tilde{Z}_n(\mathcal{K}_n^G(i)), \left( \text{span} \left( \tilde{Z}_n(\mathcal{K}_n^G(1)), \ldots, \tilde{Z}_n(\mathcal{K}_n^G(i-1)) \right) \right) \right\|_2 \geq \left( \frac{2}{3} \right)^{\mu L}
\]

Consequently, the Gram-Schmidt orthogonalization process can be successfully performed on the $l$ vectors in $\tilde{Z}_n(\mathcal{K}_n^G)$.

Proof. First we note that from Lemma 3.20, the $l$ vectors in $\tilde{Y}_n(\mathcal{K}_n^G)$ obtained after running Algorithm 2 once and Algorithm 3 at most $l$ times, satisfy Properties P 3.14 through P 3.16. Now, from (17) it is easy to see that for each $i = 1, \ldots, l$,
\[
\left\| \tilde{Z}_n(\mathcal{K}_n^G(i)) \right\|_2 = \sqrt{\left\| \tilde{Y}_n(\mathcal{K}_n^G(i)) \right\|_2^2 + \frac{1}{2} \left\| \tilde{Y}_n(\mathcal{K}_n^G(i)) \right\|_2^4}
\]

From Property P 3.15, we know that $\left\| \tilde{Y}_n(\mathcal{K}_n^G(i)) \right\|_2 \leq 1$. Using this, we get
\[
\left\| \tilde{Y}_n(\mathcal{K}_n^G(i)) \right\|_2 \leq \left\| \tilde{Z}_n(\mathcal{K}_n^G(i)) \right\|_2 \leq \sqrt{\frac{3}{2}} \left\| \tilde{Y}_n(\mathcal{K}_n^G(i)) \right\|_2
\]

(178)

Since (174) in Property P 3.16 holds, we get from (178) that for each $i = 1, \ldots, l$,
\[
\left\| \tilde{Z}_n(\mathcal{K}_n^G(i)) \right\|_2 \geq \left\| \tilde{Y}_n(\mathcal{K}_n^G(i)) \right\|_2 \geq \mu L = \left( \frac{3}{2} \right) \left( \frac{2}{3} \right) \mu L > 0
\]

Next, for each $i = 2, \ldots, l$, we get using the definition of the orthogonal projection operator, that
\[
\left\| \Pi \left( \tilde{Z}_n(\mathcal{K}_n^G(i)), \left( \text{span} \left\{ \tilde{Z}_n(\mathcal{K}_n^G(j)) : j = 1, \ldots, i-1 \right\} \right) \right) \right\|_2 = \left\| \tilde{Z}_n(\mathcal{K}_n^G(i)) - \Pi \left( \tilde{Z}_n(\mathcal{K}_n^G(i)), \text{span} \left\{ \tilde{Z}_n(\mathcal{K}_n^G(j)) : j = 1, \ldots, i-1 \right\} \right) \right\|_2 = \min \left\{ \left\| \tilde{Z}_n(\mathcal{K}_n^G(i)) - v \right\|_2 : v \in \text{span} \left\{ \tilde{Z}_n(\mathcal{K}_n^G(j)) : j = 1, \ldots, i-1 \right\} \right\}
\]

Clearly any $v \in \text{span} \left\{ \tilde{Z}_n(\mathcal{K}_n^G(j)) : j = 1, \ldots, i-1 \right\}$ can be written as $v = \sum_{j=1}^{i-1} \alpha_j \tilde{Z}_n(\mathcal{K}_n^G(j))$ for some $\alpha_1, \ldots, \alpha_{i-1} \in \mathbb{R}$. Therefore,
\[
\min \left\{ \left\| \tilde{Z}_n(\mathcal{K}_n^G(i)) - v \right\|_2 : v \in \text{span} \left\{ \tilde{Z}_n(\mathcal{K}_n^G(j)) : j = 1, \ldots, i-1 \right\} \right\} = \min_{\alpha_1, \ldots, \alpha_{i-1} \in \mathbb{R}} \left\| \tilde{Z}_n(\mathcal{K}_n^G(i)) - \sum_{j=1}^{i-1} \alpha_j \tilde{Z}_n(\mathcal{K}_n^G(j)) \right\|_2
\]

Further, using the definition of $\tilde{Z}_n(\mathcal{K}_n^G(i))$ for $i = 1, \ldots, l$, we get that for any $\alpha_1, \ldots, \alpha_{i-1} \in \mathbb{R}$,
\[
\left\| \tilde{Z}_n(\mathcal{K}_n^G(i)) - \sum_{j=1}^{i-1} \alpha_j \tilde{Z}_n(\mathcal{K}_n^G(j)) \right\|_2 = \left\| \tilde{Y}_n(\mathcal{K}_n^G(i)) - \sum_{j=1}^{i-1} \left\{ \alpha_j \tilde{Y}_n(\mathcal{K}_n^G(j)) \right\} \right\|_2 
\]

\[
\geq \left\| \tilde{Y}_n(\mathcal{K}_n^G(i)) - \sum_{j=1}^{i-1} \alpha_j \tilde{Y}_n(\mathcal{K}_n^G(j)) \right\|_2
\]
Consequently, using Lemma 3.12, we get that
\[ \left\| \Pi \left( \hat{Z}_n(\mathcal{K}_n(i)), \left( \text{span} \left\{ \hat{Z}_n(\mathcal{K}_n(j)) : j = 1, \ldots, i-1 \right\} \right)^{\perp} \right\|_2 \]
\[ = \min_{\alpha_1, \ldots, \alpha_{i-1} \in \mathbb{R}} \left\| \hat{Z}_n(\mathcal{K}_n(i)) - \sum_{j=1}^{i-1} \alpha_j \hat{Z}_n(\mathcal{K}_n(j)) \right\|_2 \]
\[ \geq \min_{\alpha_1, \ldots, \alpha_{i-1} \in \mathbb{R}} \left\| \hat{Y}_n(\mathcal{K}_n(i)) - \sum_{j=1}^{i-1} \alpha_j \hat{Y}_n(\mathcal{K}_n(j)) \right\|_2 \]
\[ = \left\| \Pi \left( \hat{Y}_n(\mathcal{K}_n(i)), \left( \text{span} \left\{ \hat{Y}_n(\mathcal{K}_n(j)) : j = 1, \ldots, i-1 \right\} \right)^{\perp} \right\|_2 \]
\[ \geq \left( \sqrt{\frac{2}{3}} \right)^{\mu_L} \left( \frac{\left\| \Pi \left( \hat{Y}_n(\mathcal{K}_n(i)), \left( \text{span} \left\{ \hat{Y}_n(\mathcal{K}_n(j)) : j = 1, \ldots, i-1 \right\} \right)^{\perp} \right\|_2}{\left\| \hat{Y}_n(\mathcal{K}_n(i)) \right\|_2} \right) \]

Finally, using (178) and (175) in Property P 3.16, we get for each \( i = 2, \ldots, l \),
\[ \left\| \Pi \left( \hat{Z}_n(\mathcal{K}_n(i)), \left( \text{span} \left\{ \hat{Z}_n(\mathcal{K}_n(j)) : j = 1, \ldots, i-1 \right\} \right)^{\perp} \right\|_2 \]
\[ \geq \left( \sqrt{\frac{2}{3}} \right)^{\mu_L} \left( \frac{\left\| \Pi \left( \hat{Y}_n(\mathcal{K}_n(i)), \left( \text{span} \left\{ \hat{Y}_n(\mathcal{K}_n(j)) : j = 1, \ldots, i-1 \right\} \right)^{\perp} \right\|_2}{\left\| \hat{Y}_n(\mathcal{K}_n(i)) \right\|_2} \right) \]

Consequently, using Lemma 3.12, we get that \( \{ \hat{Z}_n(\mathcal{K}_n(1)), \ldots, \hat{Z}_n(\mathcal{K}_n(l)) \} \subset \mathbb{R}^p \) is a linearly independent set and therefore, we can perform Gram-Schmidt orthogonalization on this set.

Thus, if we set the first \( l \) design points \( \{ x_n + y_n^i : i = 1, \ldots, l \} \) to be those in \( \mathcal{K}_n(\mathcal{K}_n) \), then from Lemma 3.21, it is clear that (160) and (161) in Assumption A 18 are satisfied for \( i = 1, \ldots, l \) with \( \mu_Q = \left( \sqrt{\frac{2}{3}} \right) \mu_L \). The advantages of such a “warm start” are obvious. First, by running Algorithm 2 we have already separated the potential inner and outer design points in \( \mathcal{K}_n \). Also, in order satisfy Assumption A 18, instead of having to pick \( p \) design points, we only have to find \( p - l \) additional points. Finally, if for any reason, we wish to stop without picking the required \( p - l \) design points we are still assured of the bound on the condition number of \( (\hat{Y}_n)^TW_n^t\hat{Y}_n^t \) as in (159).

Motivated by Lemma 3.21, we provide an algorithm next that performs the Gram-Schmidt orthogonalization process on the scaled regression vectors corresponding to the points in \( \mathcal{K}_n(\mathcal{K}_n) \) and searches the set \( \mathcal{K}_n(\mathcal{K}_n) \) for more points that can be used to satisfy Assumption A 18.

**Algorithm 5.** Let us assume that we have knowledge of the design region radius \( \delta_n \), the sets \( \mathcal{K}_n, \tilde{\mathcal{Y}}_n, \hat{Z}_n, \overline{\mathcal{N}}, \mathcal{F}_n \) and \( \overline{\mathcal{F}}_n \), the index sets \( \mathcal{K}_n^o, \mathcal{K}_n^o, \hat{K}_n^o \) such that Properties P 3.14 through P 3.16 are satisfied. Further, we assume that \( \overline{\mathcal{M}}_n^o = l \). Initialize the set \( \overline{\mathcal{V}}_n = \emptyset \). Choose a constant \( \mu_Q1 \in (0, 1] \).

**Step 1:** Repeat the following steps for each \( i = 1, \ldots, \overline{\mathcal{M}}_n^o \).
• If \( i = 1 \), then set \( v_n \leftarrow \hat{Z}_n(\mathcal{K}_n^G(i)) \).

• If \( i \geq 2 \), then set

\[
v_n \leftarrow \hat{Z}_n(\mathcal{K}_n^G(i)) - \sum_{j=1}^{i-1} \left( \hat{Z}_n(\mathcal{K}_n^G(i))^T \hat{V}_n(j) \right) \hat{V}_n(j)
\]

• Insert \( (v_n/\|v_n\|_2) \) into the set \( \hat{V}_n \) at position \( i \).

**Step 2:** Repeat the following steps for \( i = 1, \ldots, M_n^G \).

• If \( M_n^G < p \) then

  - Set

  \[
v_n \leftarrow \hat{Z}_n(\mathcal{K}_n^G(i)) - \frac{M_n^G}{\|v_n\|_2} \left( \hat{Z}_n(\mathcal{K}_n^G(i))^T \hat{V}_n(j) \right) \hat{V}_n(j)
\]

  - If \( \left\| \hat{Z}_n(\mathcal{K}_n^G(i)) \right\|_2 \geq (\sqrt{3/2})\mu Q_1 \) and \( \|v_n\|_2 \geq \mu Q_1 \left\| \hat{Z}_n(\mathcal{K}_n^G(i)) \right\|_2 \) then

    * Set \( M_n^G \leftarrow M_n^G + 1 \).
    * Insert \( \mathcal{K}_n^G(i) \) into the set \( \mathcal{K}_n^G \) at position \( M_n^G \).
    * Insert \( (v_n/\|v_n\|_2) \) into the set \( \hat{V}_n \) at the position \( M_n^G \).

  - Otherwise if \( \left\| \hat{Z}_n(\mathcal{K}_n^G(i)) \right\|_2 < (\sqrt{3/2})\mu Q_1 \) or \( \|v_n\|_2 < \mu Q_1 \left\| \hat{Z}_n(\mathcal{K}_n^G(i)) \right\|_2 \), then set \( M_{tmp} \leftarrow M_{tmp} + 1 \) and insert \( \mathcal{K}_n^G(i) \) into the set \( \mathcal{K}_{tmp}^G \) at position \( M_{tmp} \).

• Otherwise if \( M_n^G = p \) then set \( M_{tmp} \leftarrow M_{tmp} + 1 \) and insert \( \mathcal{K}_n^G(i) \) into the set \( \mathcal{K}_{tmp}^G \) at position \( M_{tmp} \).

**Step 3:** Set \( \mathcal{K}_n^G \leftarrow \mathcal{K}_{tmp}^G \) and \( M_n^G \leftarrow M_{tmp}^G \).

We know from Lemma 3.21 that we can perform Gram-Schmidt orthogonalization on the vectors in \( \hat{Z}_n(\mathcal{K}_n^G) \). Step 1 of Algorithm 5 does exactly this. Step 2 checks each scaled regression vector \( \hat{Z}_n(\mathcal{K}_n^G(i)) \) for \( i = 1, \ldots, M_n^G \), to see if it satisfies

\[
\left\| \hat{Z}_n(\mathcal{K}_n^G(i)) \right\|_2 \geq \left( \sqrt{\frac{3}{2}} \right) \mu Q_1 \quad \text{and} \quad \left\| \Pi \left( \hat{Z}_n(\mathcal{K}_n^G(i)), \text{span} \left\{ \hat{Z}_n(\mathcal{K}_n^G(i)) \right\} \right) \right\|_2 \geq \mu Q_1
\]

If it does then the index \( \mathcal{K}_n^G(i) \) is moved to \( \mathcal{K}_n^G \) and the set \( \hat{V}_n \) containing the Gram-Schmidt vectors corresponding to those in \( \hat{Z}_n(\mathcal{K}_n^G) \) is updated appropriately. Thus, it is easily seen that after the execution of Algorithm 5, Properties P 3.14 and P 3.15 continue to hold. Further from Lemma 3.21 and Steps 1 and 2 of Algorithm 5 it can be seen that, the sets \( \hat{Z}_n(\mathcal{K}_n^G) \) and \( \hat{V}_n \) satisfy the following property (analogous to Property P 3.16) for \( \mu Q = \min\{\mu Q_1, (\sqrt{2/3})\mu L\} \).

**P 3.17.** 1. The vectors in the set \( \hat{V}_n \) are the Gram-Schmidt orthonormal vectors corresponding to those in \( \hat{Z}_n(\mathcal{K}_n^G) \). That is,

\[
\hat{V}_n(1) = \frac{\hat{Z}_n(\mathcal{K}_n^G(1))}{\left\| \hat{Z}_n(\mathcal{K}_n^G(1)) \right\|_2}
\]
and for \( i = 2, \ldots, N_n^g \),

\[
\overline{V}_n(i) = \frac{\hat{Z}_n(K_n^g(i)) - \sum_{j=1}^{i-1} \left( \hat{Z}_n(K_n^g(i))V_n(j) \right) V_n(j)}{\| \hat{Z}_n(K_n^g(i)) - \sum_{j=1}^{i-1} \left( \hat{Z}_n(K_n^g(i))V_n(j) \right) V_n(j) \|^2_2} \tag{181}
\]

2. The \( M_n^g \) vectors in \( \hat{Z}_n(K_n^g) \) satisfy for each \( i = 1, \ldots, M_n^g \),

\[
\left\| \hat{Z}_n(K_n^g(i)) \right\|_2 \geq \left( \sqrt{\frac{3}{2}} \right) \mu_Q \tag{182}
\]

and from Property P 3.13, we get that for \( i = 2, \ldots, M_n^g \),

\[
\left\| \prod \left( \hat{Z}_n(K_n^g(i)), \left( \text{span} \{ \hat{Z}_n(K_n^g\{1, \ldots, i-1\}\} \} \right) \right) \right\|_2
\]

\[
\left\| \hat{Z}_n(K_n^g(i)) \right\|_2
\]

\[
= \frac{\left\| \hat{Z}_n(K_n^g(i)) - \sum_{j=1}^{i-1} \left( \hat{Z}_n(K_n^g(i))V_n(j) \right) V_n(j) \right\|_2}{\left\| \hat{Z}_n(K_n^g(i)) \right\|_2} \geq \mu_Q \tag{183}
\]

From Property P 3.17, it is clear that upon termination of Algorithm 5, we can set the first \( M_n^g \) design points to be those in \( \overline{X}_n(K_n^g) \) and satisfy (160) and (161) in Assumption A 18 with \( \mu_Q = \min(\mu_{Q1}, (\sqrt{2/3})\mu_L) \). However, in order to satisfy Assumption A 18, we require \( p \) inner design points satisfying (160) and (161), and it is clearly possible that after the execution of Algorithm 5, we get \( M_n^g < p \). Accordingly we provide next, an algorithm analogous to Algorithm 3, that adds an appropriate point into \( X_n \) and increments \( M_n^g \) by one such that Properties P 3.14 through Property P 3.17 continue to hold.

Let us recall the operation of Algorithm 3. There, if \( \text{dim}(\text{span}\{\hat{Y}_n(K_n^g)\}) = M_n^g < t \), we merely found a unit vector \( \hat{y}_n \in \text{span}\{\hat{Y}_n(K_n^g)\} \) and inserted \( x_n + \delta_n \hat{y}_n \) to \( X_n \). Now, if \( \text{dim}(\text{span}\{\hat{Z}_n(K_n^g)\}) = M_n^g < p \), we could, in the same manner, find a unit vector \( \hat{w}_n \in \text{span}\{\hat{Z}_n(K_n^g)\} \). However, in this case, in order to be able to insert a point to \( X_n, \hat{w}_n \), has to have a very specific form. That is, we can add a point \( x_n + \delta_n \hat{y}_n \) to \( X_n \) only if \( \hat{w}_n \in \text{span}\{\hat{Z}_n(K_n^g)\} \) is such that

\[
\hat{w}_n = \begin{pmatrix} \hat{y}_n \\ (\hat{y}_n)^Q \end{pmatrix}
\]

The following two lemmas show that even though we cannot pick any arbitrary \( \hat{w}_n \in \text{span}\{\hat{Z}_n(K_n^g)\} \), there exists an appropriate vector \( \hat{z}_n = \begin{pmatrix} \hat{y}_n \\ (\hat{y}_n)^Q \end{pmatrix} \) sufficiently close to \( \hat{w}_n \), such that \( x_n + \delta_n \hat{y}_n \) can be inserted into \( X_n \).

**Lemma 3.22.** Consider any \( \hat{w}_1 \in \mathbb{R}^p \) with \( \| \hat{w}_1 \|_2 = 1 \). Then, there exists \( \hat{z} \in \mathbb{R}^p \) given by \( \hat{z} = \begin{pmatrix} \hat{y} \\ \hat{y}^Q \end{pmatrix} \) for some \( \hat{y} \in \mathbb{R}^l \), such that \( \| \hat{y} \|_2 \leq 1 \) and

\[
| z^T \hat{w}_1 | = \left| \left( \hat{y}^T (\hat{y}^Q)^T \right) \hat{w}_1 \right| \geq \frac{1}{2\sqrt{p}} \tag{184}
\]
Further, given any such $\tilde{z} \in \mathbb{R}^p$ and a subspace $\mathcal{M} \subset \mathbb{R}^p$ such that $\tilde{w}_1 \in \mathcal{M}^\perp$, we have
\[
\|\tilde{z}\|_2 \geq \left(\sqrt{\frac{3}{2}}\right) \left(\frac{1}{\sqrt{6p}}\right) \quad \text{and} \quad \frac{\|\Pi (\tilde{z}, \mathcal{M}^\perp)\|_2}{\|\tilde{z}\|_2} \geq \frac{1}{\sqrt{6p}}
\]

**Proof.** Define the vector $\tilde{w} \in \mathbb{R}^p$ as
\[
\tilde{w} := \frac{\tilde{w}_1}{\|\tilde{w}_1\|_\infty}
\]
Then clearly, we have $\|\tilde{w}\|_\infty = 1$. Now, let the components of $\tilde{w} \in \mathbb{R}^p$ be as follows,
\[
\tilde{w}^T = \left(u_1, \ldots, u_i, v_{12}, v_{13}, v_{23}, \ldots, v_{11i}, \ldots, v_{(i-1)i}, v_{ii}, \ldots, v_i\right)
\]
Then, since $\|\tilde{w}\|_\infty = 1$, we get that one of the following cases must hold

(a) $|u_{j^*}| = 1$ for some $j^* \in \{1, \ldots, l\}$.

(b) $|v_{j^*j^*}| = 1$ for some $j^* \in \{1, \ldots, l\}$.

(c) $|v_{j^*k^*}| = 1$ for some $j^*, k^* \in \{1, \ldots, l\}$ where $j^* \neq k^*$.

Let for any $\tilde{y} \in \mathbb{R}^l$, the components of $\tilde{y}$ be given as $\tilde{y}^T = \left(\tilde{y}_1, \ldots, \tilde{y}_l\right)$. Then, from (16), we get that
\[
\left(\tilde{y}^T \ (\tilde{y}^2)^T\right) = \left(\tilde{y}_1, \ldots, \tilde{y}_i, \tilde{y}_1\tilde{y}_2, \tilde{y}_1\tilde{y}_3, \ldots, \tilde{y}_1\tilde{y}_l, \ldots, \tilde{y}_{l-1}\tilde{y}_l, \frac{1}{\sqrt{2}}\tilde{y}_1^2, \ldots, \frac{1}{\sqrt{2}}\tilde{y}_l^2\right)
\]
Thus, for any $\tilde{y} \in \mathbb{R}^l$, we get that
\[
\left|\left(\tilde{y}^T \ (\tilde{y}^2)^T\right)\tilde{w}\right| = \left\{\sum_{j=1}^l \left(u_j\tilde{y}_j + \frac{1}{\sqrt{2}}v_{jj}\tilde{y}_j^2\right)\right\} + \left\{\sum_{k=2}^l \sum_{j=1}^{k-1} v_{jk}\tilde{y}_j\tilde{y}_k\right\}
\]
Also, we will use the following identity repeatedly in our proof.
\[
\max\{|\alpha_j| : j = 1, \ldots, N\} \geq \sqrt{\frac{\sum_{j=1}^N |\alpha_j|^2}{N}} \quad \text{for any } \alpha_1, \ldots, \alpha_N \in \mathbb{R}
\]
Now, let use consider the three cases, listed earlier in order.

(a) Suppose that $|u_{j^*}| = 1$ for some $j^* \in \{1, \ldots, l\}$. Then consider the following two points $\tilde{y}^1, \tilde{y}^2 \in \mathbb{R}^l$.
\[
(\tilde{y}^1)^T = \left(\tilde{y}_1^1, \ldots, \tilde{y}_l^1\right) \quad \text{where} \quad \tilde{y}_{j^*}^1 = 1 \quad \text{and} \quad \tilde{y}_j^1 = 0 \quad \text{for} \quad j \neq j^*
\]
\[
(\tilde{y}^2)^T = \left(\tilde{y}_1^2, \ldots, \tilde{y}_l^2\right) \quad \text{where} \quad \tilde{y}_{j^*}^2 = -1 \quad \text{and} \quad \tilde{y}_j^2 = 0 \quad \text{for} \quad j \neq j^*
\]
Clearly, $\|\tilde{y}\|_2 = \|\tilde{y}^2\|_2 = 1$. Now, we get using the fact that $|u_{j^*}| = 1$ that,

$$\max \left\{ \left\| \left( \tilde{y}_1^T \ (\tilde{y}_1^Q)^T \right) \tilde{w} \right\| , \left\| \left( \tilde{y}_2^T \ (\tilde{y}_2^Q)^T \right) \tilde{w} \right\| \right\} = \max \left\{ \left| u_{j^*} + \frac{v_{j^*}^2}{\sqrt{2}} \right| , \left| -u_{j^*} + \frac{v_{j^*}^2}{\sqrt{2}} \right| \right\} \geq \sqrt{|u_{j^*}|^2 + \frac{v_{j^*}^2}{2}} \quad \text{(using (185))}$$

(b) Suppose that $|v_{j^*,j^*}| = 1$ for some $j^* \in \{1, \ldots, l\}$. Then again we consider the points $\tilde{y}_1, \tilde{y}_2 \in \mathbb{R}^l$ as defined in (186). Using the fact that $|v_{j^*,j^*}| = 1$, we get

$$\max \left\{ \left\| \left( \tilde{y}_1^T \ (\tilde{y}_1^Q)^T \right) \tilde{w} \right\| , \left\| \left( \tilde{y}_2^T \ (\tilde{y}_2^Q)^T \right) \tilde{w} \right\| \right\} = \max \left\{ \left| u_{j^*} + \frac{v_{j^*}^2}{\sqrt{2}} \right| , \left| -u_{j^*} + \frac{v_{j^*}^2}{\sqrt{2}} \right| \right\} = \sqrt{|u_{j^*}|^2 + \frac{v_{j^*}^2}{2}} \quad \text{(using (185))}$$

(c) Finally suppose that $|v_{j^*,k^*}| = 1$ for some $j^*, k^* \in \{1, \ldots, l\}$ where $j^* \neq k^*$. Then we consider the following four points.

$$(\tilde{y}_1^1)^T = \left( \tilde{y}_1^1 \ldots \tilde{y}_1^1 \right) \quad \text{where} \quad \tilde{y}_1^{j^*} = \frac{1}{\sqrt{2}}, \quad \tilde{y}_1^{k^*} = \frac{1}{\sqrt{2}} \quad \text{and} \quad \tilde{y}_1^j = 0 \quad \text{for} \quad j \neq j^*, k^*$$

$$(\tilde{y}_2^1)^T = \left( \tilde{y}_2^1 \ldots \tilde{y}_2^1 \right) \quad \text{where} \quad \tilde{y}_2^{j^*} = -\frac{1}{\sqrt{2}}, \quad \tilde{y}_2^{k^*} = -\frac{1}{\sqrt{2}} \quad \text{and} \quad \tilde{y}_2^j = 0 \quad \text{for} \quad j \neq j^*, k^*$$

$$(\tilde{y}_3^1)^T = \left( \tilde{y}_3^1 \ldots \tilde{y}_3^1 \right) \quad \text{where} \quad \tilde{y}_3^{j^*} = \frac{1}{\sqrt{2}}, \quad \tilde{y}_3^{k^*} = \frac{1}{\sqrt{2}} \quad \text{and} \quad \tilde{y}_3^j = 0 \quad \text{for} \quad j \neq j^*, k^*$$

$$(\tilde{y}_4^1)^T = \left( \tilde{y}_4^1 \ldots \tilde{y}_4^1 \right) \quad \text{where} \quad \tilde{y}_4^{j^*} = -\frac{1}{\sqrt{2}}, \quad \tilde{y}_4^{k^*} = -\frac{1}{\sqrt{2}} \quad \text{and} \quad \tilde{y}_4^j = 0 \quad \text{for} \quad j \neq j^*, k^*$$

Again, it is easily seen that $\|\tilde{y}_1\|_2 = \|\tilde{y}_2\|_2 = \|\tilde{y}_3\|_2 = \|\tilde{y}_4\|_2 = 1$. Now, using the fact that $|v_{j^*,k^*}| = 1$, we get

$$\max \left\{ \left\| \left( \tilde{y}_i^T \ (\tilde{y}_i^Q)^T \right) \tilde{w} \right\| : i = 1, \ldots, 4 \right\} = \max \left\{ \left( \frac{v_{j^*,j^*}^2 + v_{k^*,k^*}^2}{2} \right) + \left( \frac{u_{j^*,k^*} + u_{k^*,j^*}}{\sqrt{2}} \right) , \left( \frac{v_{j^*,j^*}^2 + v_{k^*,k^*}^2}{2} \right) + \left( \frac{u_{j^*,k^*} - u_{k^*,j^*}}{\sqrt{2}} \right) , \left( \frac{v_{j^*,j^*}^2 + v_{j^*,k^*}^2}{2} \right) - \left( \frac{u_{j^*,k^*}^2}{\sqrt{2}} \right) , \left( \frac{v_{j^*,j^*}^2 + v_{k^*,k^*}^2}{2} \right) - \left( \frac{u_{j^*,k^*}^2}{\sqrt{2}} \right) \right\} \geq \frac{|v_{j^*,k^*}|}{2} = \frac{1}{2} \quad \text{(using (185))}$$

Therefore, we see that in all three cases, there exists a vector $\tilde{\xi} = \begin{pmatrix} \tilde{y} \\ \tilde{y}^Q \end{pmatrix}$, such that $\|\tilde{\xi}\|_2 \leq 1$ and

$$\|\tilde{\xi}^T \tilde{w}\| = \left\| \left( \tilde{y}^T \ (\tilde{y}^Q)^T \right) \tilde{w} \right\| \geq \frac{1}{2}$$
Thus, since $\|\tilde{w}_1\|_2 = 1$, 
\[
|(\tilde{z})^T \tilde{w}_1| = \|\tilde{w}_1\|_\infty |(\tilde{z})^T \tilde{w}_1| \geq \frac{\|\tilde{w}_1\|_\infty}{2} \geq \frac{1}{2\sqrt{p}}
\]

Next since $\tilde{w}_1 \in M^\perp$ and $\|\tilde{w}_1\|_2 = 1$, let us extend it an orthonormal basis for $M^\perp$ given by \(\{\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_k\}\) \(\subset M^\perp\). Let $\tilde{z} \in \mathbb{R}^p$ by such that 
\[
\tilde{z} = \begin{pmatrix} \tilde{y} \\ \tilde{y}Q \end{pmatrix} \quad \text{with} \quad \tilde{y} \in \mathbb{R}^l \quad \text{and} \quad \|\tilde{y}\|_2 \leq 1
\]

Now, we know from Property P 3.8 of the projection operation, that 
\[
\Pi(\tilde{z}, M^\perp) = \sum_{j=1}^{k} ((\tilde{z})^T \tilde{w}_j) \tilde{w}_j
\]

Therefore, we get 
\[
\frac{\|\Pi(\tilde{z}, M^\perp)\|_2}{\|\tilde{z}\|_2} = \sqrt{\sum_{j=1}^{k} ((\tilde{z})^T \tilde{w}_j)^2} \geq |(\tilde{z})^T \tilde{w}_1| \geq \frac{1}{2\sqrt{p}}
\]

Therefore, using Property P 3.6, we get that 
\[
\|\tilde{z}\|_2 \geq \frac{\|\Pi(\tilde{z}, M^\perp)\|_2}{\|\tilde{z}\|_2} \geq \frac{1}{2\sqrt{p}} \geq \left(\frac{\sqrt{3}}{2}\right) \left(\frac{1}{\sqrt{6p}}\right)
\]

Further since $\|\tilde{y}\|_2 \leq 1$ get from (17) that 
\[
\|\tilde{z}\|_2 = \sqrt{\|\tilde{y}\|_2^2 + \frac{1}{2} \|\tilde{y}\|_2^4} \leq \sqrt{\frac{3}{2}}
\]

Consequently, we get 
\[
\frac{\|\Pi(\tilde{z}, M^\perp)\|_2}{\|\tilde{z}\|_2} \geq \frac{\frac{1}{\sqrt{6p}}}{\left(\sqrt{\frac{3}{2}}\right)} = \frac{1}{\sqrt{6p}}
\]

Lemma 3.22 motivates the following algorithm that adds a single appropriate point to $\mathcal{X}_n$.

**Algorithm 6.** Let us assume that we have knowledge of the design region radius $\delta_n$, the sets $\mathcal{X}_n$, $\tilde{y}_n$, $\tilde{z}_n$, $\mathcal{N}_n$, $\mathcal{F}_n$ and $\mathcal{B}_n$ and $\mathcal{V}_n$ and the index sets $\mathcal{K}^G_n$, $\mathcal{K}^B_n$ and $\mathcal{K}^O_n$ such that Properties P 3.14 through P 3.17 are satisfied. Let $l \leq \overline{M}^G_n < p$ and let $\{e^1, \ldots, e^p\}$ denote the standard basis for $\mathbb{R}^p$

**Step 1:** Find an $i^* \in \{1, \ldots, p\}$ such that 
\[
v_n = e^{i^*} - \sum_{j=1}^{\overline{M}^G_n} \left((e^{i^*})^T \mathcal{V}_n(j)\right) \mathcal{V}_n(j)
\]
satisfies $\|v_n\|_2 > 0$. 

85
Step 2: Compute $\tilde{z}_n = \begin{pmatrix} \tilde{y}_n \\ (\tilde{y}_n)^Q \end{pmatrix} \in \mathbb{R}^p$ with $\tilde{y}_n \in \mathbb{R}^l$ such that $\|\tilde{y}_n\|_2 \leq 1$ and
\[
\left\| \tilde{z}_n ^T \left( \frac{v_n}{\|v_n\|_2} \right) \right\| \geq \frac{1}{2\sqrt{p}} \tag{187}
\]

Step 3: Set $\mathcal{M}_n \leftarrow \mathcal{M}_n + 1$.

Step 4: Insert $\tilde{y}_n$ into $\tilde{Y}_n$, $\tilde{z}_n$ into $\tilde{Z}_n$ and $x_n + \delta_n \tilde{y}_n$ into $\mathcal{X}_n$ all at position $\mathcal{M}_n$.

Step 5: Insert the value $0$ into $\mathcal{F}_n$, $\Phi_n$ and $\mathcal{N}_n$ all at position $\mathcal{M}_n$.

Step 6: Set $\mathcal{M}_G^n \leftarrow \mathcal{M}_G^n + 1$ and insert $\mathcal{M}_n$ into the set $\mathcal{K}_G^n$ at position $\mathcal{M}_G^n$.

Step 7: Set $v_n \leftarrow \tilde{z}_n - \sum_{j=1}^{\mathcal{M}_n-1} (\tilde{z}_n ^T \mathcal{V}_n(j)) \mathcal{V}_n(j)$

Step 8: Insert $(v_n/\|v_n\|_2)$ into $\mathcal{V}_n$ at position $\mathcal{M}_G^n$.

Note: In Step 2, we could potentially calculate the vector $\tilde{z}_n = \begin{pmatrix} \tilde{y}_n \\ (\tilde{y}_n)^Q \end{pmatrix}$ as
\[
\tilde{z}_n \in \arg \max \left\{ \left\| \tilde{z}_n ^T \left( \frac{v_n}{\|v_n\|_2} \right) \right\| : \tilde{z} = \begin{pmatrix} \tilde{y} \\ \tilde{y}^Q \end{pmatrix} \text{ where } \tilde{y} \in \mathbb{R}^l \text{ and } \|\tilde{y}\|_2 \leq 1 \right\} \tag{188}
\]
If $\tilde{z}_n$ is chosen as in (188), then clearly, from Lemma 3.22, (187) is satisfied. However, to find $\tilde{z}_n$ as in (188), it is easily seen that we need to solve two nonlinearly constrained quadratic optimization problems which could be computationally expensive. Alternatively, we can find an appropriate vector $\tilde{z}_n$ satisfying (187) by a simple enumeration as shown in the proof of Lemma 3.22.

The following lemma shows that as long as $\mathcal{M}_n^G < p$ before the execution of Algorithm 6, the algorithm successfully inserts an appropriate point into $\mathcal{X}_n$ such that Properties P 3.14 and P 3.15 continue to hold and Property P 3.17 continues to hold.

Lemma 3.23. Suppose that the sets $\mathcal{X}_n$, $\tilde{Y}_n$, $\tilde{Z}_n$, $\mathcal{N}_n$, $\mathcal{F}_n$, $\Phi_n$ and $\mathcal{V}_n$ and the index sets $\mathcal{K}_G^n$, $\mathcal{K}_B^n$ and $\mathcal{K}_O^n$ satisfy Properties P 3.14 and P 3.15 and Property P 3.17 for $\mu_Q^G = \min\{((\sqrt{2/3})\mu_L, \mu_{Q1}, (1/\sqrt{6p})\}$ and we have $\mathcal{M}_n^G < p$ before the execution of Algorithm 6. Then, the algorithm is well defined and upon its termination, an appropriate point is added to $\mathcal{X}_n$ and $\mathcal{M}_n^G$ is incremented by one such that Properties P 3.14 and P 3.15 continue to hold and Property P 3.17 holds for $\mu_Q = \min\{((\sqrt{2/3})\mu_L, \mu_{Q1}, (1/\sqrt{6p})\}$.

Proof. First we show that as long as $\mathcal{M}_n^G < p$ before the execution of Algorithm 6, all its steps are well defined. Let us begin with Step 1. From our assumptions we know that $\mathcal{M}_n^G < p$ before the execution of the
algorithm. Since Property P 3.17 holds before the execution of the algorithm, we know from the first part of this property that $\mathcal{V}_n$ contains a set of $\mathcal{M}_n^G < p$ Gram-Schmidt vectors corresponding to those in $\tilde{Z}_n(\mathcal{K}_n^G)$. Therefore, $\dim(\text{span}(\mathcal{V}_n)) = \dim\left(\text{span}\left\{\tilde{Z}_n(\mathcal{K}_n^G)\right\}\right) = \mathcal{M}_n^G < p$. Since $\dim(\text{span}\{e_1, \ldots, e_p\}) = p$, clearly there exists $i^* \in \{1, \ldots, p\}$ such that $e^{i^*} \notin \text{span}(\mathcal{V}_n)$, i.e., such that

$$\|v_n\|_2 = \left\|e^{i^*} - \sum_{j=1}^{\mathcal{M}_n^G} \left((e^{i^*})^T \mathcal{V}_n(j)\right) \mathcal{V}_n(j)\right\|_2 = \left\|\left(e^{i^*}, \left(\text{span}(\tilde{Z}_n(\mathcal{K}_n^G))\right) \right)^\perp\right\|_2 > 0$$

Therefore, Step 1 is well defined. We know from Lemma 3.22, that there exists $\tilde{y}_n \in \mathbb{R}^1$ with $\|\tilde{y}_n\|_2 \leq 1$ such that (187) holds. Therefore, Step 2 in Algorithm 6 is well defined. It is trivial to see that the rest of the steps in the algorithm are well defined and consequently that the algorithm itself is well defined.

Next we show that if Properties P 3.14, P 3.15 and P 3.17 were satisfied before the execution of Algorithm 6, they continue to be satisfied after its execution. It is clear that upon termination of Algorithm 6, $\mathcal{M}_n$ is incremented by one and $\tilde{y}_n(\mathcal{M}_n) = \tilde{y}_n$, $\tilde{Z}_n(\mathcal{M}_n) = \left(\tilde{y}_n\right)^Q$ and $\mathcal{X}_n(\mathcal{M}_n) = x_n + \delta_n \tilde{y}_n$. Since no other changes are made to $\mathcal{X}_n$, $\tilde{y}_n$ and $\tilde{Z}_n$ and Property P 3.14 we assumed to hold before the algorithms execution, it is clear that this property continues to hold. Further, we also know that Algorithm 6 increments $\mathcal{M}_n^G$ by one and sets $\mathcal{K}_n^G(\mathcal{M}_n^G) = \mathcal{M}_n$. It is easily seen that the incremented value of $\mathcal{M}_n$ cannot lie in either $\mathcal{K}_n^a$ or $\mathcal{K}_n^G$. Thus, since a new entry($\mathcal{M}_n$) is inserted only into $\mathcal{K}_n^a$ and no changes are made to $\mathcal{K}_n^a$ or $\mathcal{K}_n^G$, the first part of Property P 3.15 continues to hold after the execution of Algorithm 6. Finally, since $\tilde{y}_n$ as defined in (187) in Step 2 satisfies $\|\tilde{y}_n\|_2 \leq 1$, it is also is clear that the second part of Property P 3.15 continues to hold.

Next, let us show that if Property P 3.17 holds before the execution of Algorithm 6 for $\mu_Q = \min\{(\sqrt{2/3})\mu_L, \mu_Q, (1/\sqrt{p})\}$ then the property holds after the execution of Algorithm 6 for the same value of $\mu_Q$. Note again that upon termination of the algorithm, the value of $\mathcal{M}_n^G$ is incremented by one with $\tilde{Z}_n(\mathcal{K}_n^G(\mathcal{M}_n^G)) = \tilde{z}_n$ and from Step 8,

$$\mathcal{V}_n(\mathcal{M}_n^G) = \frac{\tilde{Z}_n(\mathcal{K}_n^G(\mathcal{M}_n^G)) - \sum_{j=1}^{\mathcal{M}_n^G-1} \left(\tilde{Z}_n(\mathcal{K}_n^G(\mathcal{M}_n^G))^T \mathcal{V}_n(j)\right) \mathcal{V}_n(j)}{\left\|\tilde{Z}_n(\mathcal{K}_n^G(\mathcal{M}_n^G)) - \sum_{j=1}^{\mathcal{M}_n^G-1} \left(\tilde{Z}_n(\mathcal{K}_n^G(\mathcal{M}_n^G))^T \mathcal{V}_n(j)\right) \mathcal{V}_n(j)\right\|_2}$$

(189)

Since the first part Property P 3.17 held before the execution of the algorithm, clearly after its execution, (180) holds and (181) holds for $i = 2, \ldots, \mathcal{M}_n^G - 1$. From (189) it is clear that (181) holds for $i = \mathcal{M}_n^G$. Therefore the first part of Property P 3.17 continues to hold after the execution of Algorithm 6.

Similarly, since the second part of Property P 3.17 was assumed to hold before the execution of Algorithm 6, we know that after its execution, (182) and (183) hold for $i = 1, \ldots, \mathcal{M}_n^G - 1$ and $\mu_Q = \min\{(\sqrt{2/3})\mu_L, \mu_Q, (1/\sqrt{p})\}$. Thus, it only remains for us to show that (182) and (183) hold for $i = \mathcal{M}_n^G$ and the same value of $\mu_Q$. Now, we know that after the algorithm is executed and $\mathcal{M}_n^G$ is incremented, the
vector $v_n$ computed in Step 1 of the algorithm satisfies

$$v_n = e^* - \sum_{j=1}^{\hat{M}_n^G-1} \left( (e^*)^T \nabla_n(j) \right) \nabla_n(j) \in \text{span}\{\nabla_n(\{1, \ldots, \hat{M}_n^G-1\})\}$$

$$= \text{span}\left\{ \tilde{z}_n \left( \mathcal{K}_n^G \left( \{1, \ldots, \hat{M}_n^G-1\} \right) \right) \right\}$$

Thus, $(v_n/\|v_n\|_2)$ used in Step 2 is a unit vector in the subspace $\text{span}\left\{ \tilde{z}_n \left( \mathcal{K}_n^G \left( \{1, \ldots, \hat{M}_n^G-1\} \right) \right) \right\}$. Since we chose $\tilde{z}_n(\mathcal{K}_n^G(\hat{M}_n^G)) = \tilde{z}_n$ in Step 2 such that (187) is satisfied, we get using Lemma 3.22 that

$$\left\| \tilde{z}_n(\mathcal{K}_n^G(\hat{M}_n^G)) \right\|_2 \geq \left( \frac{\sqrt{3}}{2} \right) \left( \frac{1}{\sqrt{6p}} \right) \geq \left( \frac{\sqrt{3}}{2} \right) \mu_Q$$

and

$$\left\| \Pi \left( \tilde{z}_n(\mathcal{K}_n^G(\hat{M}_n^G)), (\text{span}\left\{ \tilde{z}_n(\mathcal{K}_n^G(\{1, \ldots, \hat{M}_n^G-1\})) \right\}) \right) \right\|_2 \geq \left( \frac{1}{\sqrt{6p}} \right) \geq \mu_Q$$

Therefore, (182) and (183) hold for $i = \hat{M}_n^G$. Thus, it is clear that Property P 3.17 holds after the execution of Algorithm 6 for $\mu_Q = \min\{((2\sqrt{3})\mu_L, \mu_{Q1}(1/\sqrt{6p}))\}$.

Finally, we show that if we have the set $\mathcal{X}_n$ (and the corresponding $\mathcal{N}_n$, $\mathcal{F}_n$ and $\mathcal{F}_n^b$) with $\hat{M}_n \leq M_{\text{max}}$ at the beginning of each $n \in \mathbb{N}$, we can use Algorithms 2, 3, 5 and 6, to find a set of design points $\{x_n + y_n^i : i = 1, \ldots, M_n\}$ such that for each $n \in \mathbb{N}$, Assumptions A 16 and 18 are satisfied.

**Algorithm 7.** Let us assume that for each $n \in \mathbb{N}$, we have the sets $\mathcal{X}_n$, $\mathcal{N}_n$, $\mathcal{F}_n$ and $\mathcal{F}_n^b$ with $\hat{M}_n \leq M_{\text{max}}$. Initialize the index sets $\mathcal{K}_n^G = \mathcal{K}_n^B = \mathcal{K}_n^O = \emptyset$ and the corresponding counters $M_n^G = M_n^B = M_n^O = 0$. Let $M_{\text{max}}^I \geq p$.

**Step 1:** Run Algorithm 2 once and Algorithm 3 repeatedly (at most $l$ times) until $\hat{M}_n^G = l$, to obtain the sets $\mathcal{Y}_n$ and $\tilde{z}_n$, and the index sets $\mathcal{X}_n^G$, $\mathcal{K}_n^G$ and $\mathcal{K}_n^O$.

**Step 2:** Run Algorithm 5 once and Algorithm 6 repeatedly (at most $p - l$ times) until $\hat{M}_n^G = p$.

**Step 3:** Set $\mathcal{K}_n^G = \mathcal{X}_n^G$ and $M_n^G \leftarrow \hat{M}_n^G$.

**Step 4:** Set the set $\mathcal{K}_n^B$ with $0 \leq M_n^B \leq \min\{\hat{M}_n^B, M_{\text{max}}^I - p\}$ elements, such that it contains a subset of the elements in $\mathcal{K}_n^B$ in some order (the order can be chosen arbitrarily). For example, we may set $\mathcal{K}_n^B \leftarrow \emptyset$ or $\mathcal{K}_n^B \leftarrow \mathcal{K}_n^B$ or $\mathcal{K}_n^B \leftarrow \{i \in \mathcal{K}_n^B : \mathcal{N}_n(\mathcal{K}_n^O(i)) \geq \alpha N_n^0\}$ for some constant $\alpha \in (0, 1]$.

**Step 5:** Set the set $\mathcal{K}_n^O$ with $0 \leq M_n^O \leq \hat{M}_n^O$ elements, such that it contains a subset of the elements in $\mathcal{K}_n^O$ in some order (the order can be chosen arbitrarily). For example, we may set $\mathcal{K}_n^O \leftarrow \emptyset$ or $\mathcal{K}_n^O \leftarrow \mathcal{X}_n$ or $\mathcal{K}_n^O \leftarrow \{i \in \mathcal{K}_n^O : \|\tilde{y}_n(\mathcal{K}_n^O(i))\|_2 \leq \alpha_3\}$ for some constant $\alpha_3 > 1$.

**Step 6:** For each $i = 1, \ldots, M_n^G$, set $x_n + y_n^i \leftarrow \mathcal{X}_n(\mathcal{K}_n^O(i))$. 

88
Step 7: If $M_n^B > 0$, then

- For each $i = M_n^G + 1, \ldots, M_n^G + M_n^B$, set $x_n + y_n^i \leftarrow \overline{X}_n(K_n^G(i - M_n^G))$.

Step 8: If $M_n^O > 0$, then

- For each $i = M_n^G + M_n^B + 1, \ldots, M_n^G + M_n^B + M_n^O$, set $x_n + y_n^i \leftarrow \overline{X}_n(K_n^O(i - M_n^G - M_n^B))$.

If Algorithm 7 is run for each $n \in \mathbb{N}$, then it is easily seen that the following statements hold.

1. The number of entries in $K_n^G$ is equal to $p$ and the number of entries in $K_n^B$ is at most $M_{\text{max}}^I - p$. Thus, since $\overline{X}_n(K_n^G) \cup \overline{X}_n(K_n^B)$ form the set of inner design points, it is clear that there exist at most $M_{\text{max}}^I$ inner design points for each $n \in \mathbb{N}$. Therefore, Assumption A 16 is satisfied for each $n \in \mathbb{N}$.

2. From Lemma 3.23, we know that Properties P 3.14, P 3.15 and P 3.17 are satisfied after the execution of Steps 1 and 2 in Algorithm 7. Thus, from Property P 3.17 we know that the $p$ vectors in $\tilde{Z}_n(K_n^G)$ satisfy (182) and (183) and from Property P 3.14, we know that these vectors are the scaled regression vectors corresponding to the points in $\overline{X}_n(K_n^G)$.

Also, we note from Steps 3 and 6 of Algorithm 7, that the points in $\overline{X}_n(K_n^G) = \overline{X}_n(K_n^G)$ are chosen to be the first $p$ design points. Therefore, from (182) and (183), we get that (160) and (161) are satisfied by the first $p$ design points and consequently Assumption A 18 is satisfied.

Thus, for each $n \in \mathbb{N}$, as long as the design points are chosen using Algorithm 7 and the weights $\{w_n^i : i = 1, \ldots, M_n^I\}$ are chosen so as to satisfy Assumption A 15, we get from (98), (105) and (162) that (163) holds, i.e. (58) in Assumption A 11 holds with $K_n^I = (K_n^I M_{\text{max}}^I / \mu Q_S p(\mu Q))$ for each $n \in \mathbb{N}$, where $\mu Q = \min\{(\sqrt{2/3}) \mu_L, \mu Q_1, (1/\sqrt{6p})\}$.

Finally, in this section, we consider (35) in Assumption A 7 and (57) in Assumption A 11. In both of these conditions, we require all the design points $\{x_n + y_n^i : i = 1, \ldots, M_n\}$ and $n \in \mathbb{N}$ to lie in the compact set $\mathcal{C} \subset \mathcal{E}$. Let us see how this can satisfied. Suppose that the following assumption holds.

A 19. There exists a compact set $\tilde{\mathcal{C}} \subset \mathcal{X}$ such that $\{x_n\}_{n \in \mathbb{N}} \subset \tilde{\mathcal{C}}$.

Then, since $\mathcal{X} \subset \mathcal{E}$, we get that $\tilde{\mathcal{C}} \subset \mathcal{E}$. Since $\tilde{\mathcal{C}}$ is compact and $\mathcal{E}$ is open, we get from we get from Lemma 2.1 in 7, that there exists $\delta^* > 0$ and a compact set

$$\mathcal{C} := \{x = x^1 + x^2 : x^1 \in \tilde{\mathcal{C}} \text{ and } \|x^2\|_2 \leq \delta^*\}$$

such that $\mathcal{C} \subset \mathcal{E}$. Now, suppose we choose our design region radii such that $\delta_n \leq \delta^*$ for each $n \in \mathbb{N}$. Then, it is clear that for each $n \in \mathbb{N}$, we have $\{x_n + y_n^i : i = 1, \ldots, M_n^I\} \subset \mathcal{C} \subset \mathcal{E}$. Further, note that in Algorithms 4 and 7, we never evaluate sample averages at new design points that lie outside the design region.
for any $n \in \mathbb{N}$. Instead, the only outer design points we include at each iteration $n$ are the points in \( \overline{\mathcal{X}}_n(M_{n}^O) \subset \mathcal{X}_n(M_{n}^O) \). Now, recall our description of the use of \( \mathcal{X}_n \) in our trust region algorithms. For $n=1$, \( \mathcal{X}_n \) would start as an empty set. As and when we evaluated sample averages at new inner design points, we insert these points into \( \mathcal{X}_n \). Then, at the end of each iteration $n$, the set \( \bar{\mathcal{X}}_{n+1} \) is made up of at most $M_{\text{max}}$ points appropriately chosen from \( \mathcal{X}_n \). Thus, in our algorithms, for any $n \in \mathbb{N}$ and $i=1,\ldots,M_{n}$, \( \mathcal{X}_n(K_{n}^O(i)) \) was an inner design point at an earlier iteration, i.e., \( \mathcal{X}_n(K_{n}^O(i)) \in \mathcal{D}_j \) for some $j \in \{1,\ldots,n-1\}$. Consequently, it is easy to see that under Assumption A 19, if for each $n \in \mathbb{N}$ we choose \( \delta_n \leq \delta^* \), use either Algorithm 4 or Algorithm 7 to pick \( \{x_n+y_n^i: i=1,\ldots,M_{n}\} \) and update the set \( \bar{\mathcal{X}}_{n+1} \) using points only from \( \mathcal{X}_n \), then (35) in Assumption A 7 and (57) in Assumption A 11 are satisfied.

### 3.2.2 Sample Sizes

The conditions imposed on the sample sizes are stated in Assumption A 5 for Theorem 3.2 and in Assumption A 10 for Theorem 3.7. In order to set the sample sizes \( \{N_{n}^0: i=1,\ldots,M_{n}\} \), we assume that we know the sample size \( N_{0}^n \) for each $n \in \mathbb{N}$ and that \( \{N_{0}^n\}_{n\in\mathbb{N}} \) is a monotonically increasing sequence with \( N_{0}^n \to 0 \) as \( n \to \infty \). In Section ??, when we describe a typical trust region algorithm that uses the regression model function, we will show how such a sequence \( \{N_{0}^n\}_{n\in\mathbb{N}} \) can be adaptively generated.

First, let us consider Assumption A 5 in Theorem 3.2 requires that

$$
\lim_{n \to \infty} \min_{i \in \{1,\ldots,M_{n}\}} N_{n}^i = \infty
$$

That is, the minimum sample size used to evaluate sample averages at the inner design points, must increase to infinity as $n \to \infty$. Recall that in the previous section, in order to satisfy (36) from Assumption A 7 in Theorem 3.2, we picked the design points using Algorithm 4. In Algorithm 4, we picked our inner design points to be those in \( \mathcal{X}_n(K_{n}^O) \cup \mathcal{X}_n(K_{n}^B) \). Accordingly, now we provide an algorithm to update the corresponding entries in \( \mathcal{N}_n \), \( \mathcal{F}_n \) and \( \mathcal{P}_n \) so that Assumption A 5 may be satisfied.

**Algorithm 8.** Let us assume that we have the sample size \( N_{n}^0 \), the sets \( \mathcal{N}_n \), \( \mathcal{F}_n \) and \( \mathcal{P}_n \), and the index sets \( K_{n}^O \) (\( M_{n}^O = l \)), \( K_{n}^B \) and \( K_{n}^O \) obtained after the execution of Algorithm 4. Initialize the temporary index set \( \mathcal{K}_{n}^{\text{tmp}} = \emptyset \) and \( M_{n}^{\text{tmp}} = 0 \) and choose a constant $\alpha \in (0,1]$. For any $a \in \mathbb{R}$, let $\lceil a \rceil$ denote the smallest integer greater than or equal to $a$.

**Step 1 :** First set \( \mathcal{K}_{n}^{\text{tmp}} \leftarrow K_{n}^O \) and \( M_{n}^{\text{tmp}} \leftarrow M_{n}^O \) and repeat the following steps for each $i = 1,\ldots,M_{n}^{\text{tmp}}$. Then set \( \mathcal{K}_{n}^{\text{tmp}} \leftarrow K_{n}^B \) and \( M_{n}^{\text{tmp}} \leftarrow M_{n}^B \) and again repeat the following steps for $i = 1,\ldots,M_{n}^{\text{tmp}}$.

- If \( N_{n}^0(\mathcal{K}_{n}^{\text{tmp}}(i)) < \alpha N_{n}^0 \), then
  - Set \( N \leftarrow \lceil \alpha N_{n}^0 \rceil \).
- If \( \mathcal{N}_n(\mathcal{K}^{\text{tmp}}(i)) > 0 \), then compute \( \hat{f}(\mathcal{X}_n(\mathcal{K}^{\text{tmp}}(i)), N) \) using \( \mathcal{F}_n(\mathcal{K}^{\text{tmp}}(i)) \) in (164) and 
  \( \sigma(\mathcal{X}_n(\mathcal{K}^{\text{tmp}}(i)), N) \) using \( \mathcal{N}_n(\mathcal{K}^{\text{tmp}}(i)), \mathcal{F}_n(\mathcal{K}^{\text{tmp}}(i)) \), \( \hat{f}(\mathcal{X}_n(\mathcal{K}^{\text{tmp}}(i)), N) \) and \( \overline{\Phi}_n(\mathcal{K}^{\text{tmp}}(i)) \) in (170).
  
- If \( \mathcal{N}_n(\mathcal{K}^{\text{tmp}}(i)) = 0 \), then compute \( \hat{f}(\mathcal{X}_n(\mathcal{K}^{\text{tmp}}(i)), N) \) and \( \sigma(\mathcal{X}_n(\mathcal{K}^{\text{tmp}}(i)), N) \) using (8) and (169) respectively.

**Step 2:** Set \( N^i_n \leftarrow \mathcal{N}_n(\mathcal{K}^0_n(i)) \) for each \( i = 1, \ldots, M^G_n \).

**Step 3:** If \( M^B_n > 0 \), set \( N^i_n \leftarrow \mathcal{N}_n(\mathcal{K}^B_n(i - M^B_n)) \) for each \( i = M^G_n + 1, \ldots, M^G_n + M^B_n \).

**Step 4:** If \( M^O_n > 0 \), set \( N^i_n \leftarrow \mathcal{N}_n(\mathcal{K}^O_n(i - M^G_n - M^B_n)) \) for each \( i = M^G_n + M^B_n + 1, \ldots, M^G_n + M^B_n + M^O_n \).

For each \( n \in \mathbb{N} \), upon execution of Algorithm 8, it is easily seen that \( N^i_n \geq \alpha N^0_n \) for each \( i = 1, \ldots, M^I_n \). Therefore, if we ensure that \( N^0_n \rightarrow \infty \) as \( n \rightarrow \infty \), then clearly Assumption A 5 is satisfied.

Next, in order to satisfy Assumption A 10, we have to set \( N^i_n = N^i_n^\# \) for each \( i = 1, \ldots, M_n \) and \( n \in \mathbb{N} \), where \( N^i_n^\# \rightarrow \infty \) as \( n \rightarrow \infty \). In the previous section, we used Algorithm 7 to pick the design points so as to satisfy (58) from Assumption A 11 in Theorem 3.7. Recall that upon execution of Algorithm 7, the inner design points were set to be those in \( \mathcal{X}_n(\mathcal{K}^O_n) \cup \mathcal{X}_n(\mathcal{K}^B_n) \), and the outer design points were those in \( \mathcal{X}_n(\mathcal{K}^G_n) \). Therefore, we will provide an algorithm next that updates the corresponding elements of \( \mathcal{N}_n, \mathcal{F}_n \) and \( \Phi_n \).

**Algorithm 9.** Let us assume that we have the sample size \( N^0_n \), the sets \( \mathcal{N}_n, \mathcal{F}_n \) and \( \Phi_n \), and the index sets \( \mathcal{K}^G_n \) (\( M^G_n = l \)), \( \mathcal{K}^B_n \) and \( \mathcal{K}^O_n \) obtained after the execution of Algorithm 7. Initialize the temporary index set \( \mathcal{K}^{\text{tmp}} = \emptyset \) and \( M^{\text{tmp}} = 0 \) and choose a constant \( \alpha \in (0, 1] \). For any \( a \in \mathbb{R} \), let \( [a] \) denote the smallest integer greater than or equal to \( a \).

**Step 1:** Set

\[
N^\#_n \leftarrow \max \left\{ \alpha N^0_n, \max \left\{ \mathcal{N}_n(k) : k \in \mathcal{K}^G_n \cup \mathcal{K}^B_n \cup \mathcal{K}^O_n \right\} \right\}
\]

**Step 2:** First set \( \mathcal{K}^{\text{tmp}} \leftarrow \mathcal{K}^G_n \) and \( M^{\text{tmp}} \leftarrow M^G_n \) and repeat the following steps for each \( i = 1, \ldots, M^{\text{tmp}} \).

Then set \( \mathcal{K}^{\text{tmp}} \leftarrow \mathcal{K}^B_n \) and \( M^{\text{tmp}} \leftarrow M^B_n \) and again repeat these steps for each \( i = 1, \ldots, M^{\text{tmp}} \).

Finally set \( \mathcal{K}^{\text{tmp}} \leftarrow \mathcal{K}^O_n \) and \( M^{\text{tmp}} \leftarrow M^O_n \) and repeat these steps for each \( i = 1, \ldots, M^{\text{tmp}} \).

- If \( N^0_n(\mathcal{K}^{\text{tmp}}(i)) < N^\#_n \), then
  - Set \( N^\#_n \leftarrow \alpha N^0_n \).
  - If \( \mathcal{N}_n(\mathcal{K}^{\text{tmp}}(i)) > 0 \), then compute \( \hat{f}(\mathcal{X}_n(\mathcal{K}^{\text{tmp}}(i)), N^\#_n) \) using \( \mathcal{F}_n(\mathcal{K}^{\text{tmp}}(i)) \) in (164) and 
    \( \sigma(\mathcal{X}_n(\mathcal{K}^{\text{tmp}}(i)), N^\#_n) \) using \( \mathcal{N}_n(\mathcal{K}^{\text{tmp}}(i)), \mathcal{F}_n(\mathcal{K}^{\text{tmp}}(i)) \), \( \hat{f}(\mathcal{X}_n(\mathcal{K}^{\text{tmp}}(i)), N^\#_n) \) and 
    \( \overline{\Phi}_n(\mathcal{K}^{\text{tmp}}(i)) \) in (170).
- If $\mathcal{N}_n(\mathcal{K}^{imp}(i)) = 0$, then compute $\hat{f}(\mathcal{N}_n(\mathcal{K}^{imp}(i)), N_n^{\#})$ and $\sigma(\mathcal{N}_n(\mathcal{K}^{imp}(i)), N_n^{\#})$ using (8) and (16) respectively.

- Set $\mathcal{N}_n(\mathcal{K}^{imp}(i)) \leftarrow N_n^{\#}$, $\mathcal{F}_n(\mathcal{K}^{imp}(i)) \leftarrow \hat{f}(\mathcal{N}_n(\mathcal{K}^{imp}(i)), N_n^{\#})$ and $\mathcal{B}_n(\mathcal{K}^{imp}(i)) \leftarrow \sigma(\mathcal{N}_n(\mathcal{K}^{imp}(i)), N_n^{\#})$.

**Step 3**: Set $N_n^i \leftarrow \mathcal{N}_n(\mathcal{K}^C(i))$ for each $i = 1, \ldots, M_n^O$.

**Step 4**: If $M_n^B > 0$, set $N_n^i \leftarrow \mathcal{N}_n(\mathcal{K}^B(i - M_n^O))$ for each $i = M_n^O + 1, \ldots, M_n^O + M_n^B$.

**Step 5**: If $M_n^O > 0$, set $N_n^i \leftarrow \mathcal{N}_n(\mathcal{K}^O(i - M_n^O - M_n^B))$ for each $i = M_n^O + M_n^B + 1, \ldots, M_n^O + M_n^B + M_n^O$.

From the definition of $N_n^{\#}$ in Step 1 of the above algorithm, it is clear that before the execution of Step 2, $\mathcal{N}_n(k) \leq N_n^{\#}$ for each $k \in \mathcal{K}^O_n \cup \mathcal{K}^B_n \cup \mathcal{K}^C_n$. Thus for each $n \in \mathbb{N}$, upon termination of the algorithm, we clearly get that $N_n^i = N_n^{\#} \geq \alpha N_n^0$ for each $i = 1, \ldots, M_n$. Consequently, since we ensure the $N_n^0 \rightarrow \infty$ as $n \rightarrow \infty$, if we run Algorithm 9 for each $n \in \mathbb{N}$, then Assumption A 10 is satisfied.

### 3.2.3 Weights

In this section we consider methods to set the weights $\{w_n^i : i = 1, \ldots, M_n\}$ for each $n \in \mathbb{N}$. Note that the weight matrices corresponding to the inner and outer design points, specified in the matrices $W_n^I$ and $W_n^O$ respectively, occur in (36) and (37) in Assumption A 7 and (58) and (59) in Assumption A 11. We already showed in Section 3.2.1, how the design points may picked for each $n \in \mathbb{N}$, so as to satisfy (36) and (58), using Assumption A 15. Here, we provide methods to pick the weights $\{w_n^i : i = 1, \ldots, M_n\}$ for each $n \in \mathbb{N}$ such that the conditions (37) in Assumption A 7, (59) in Assumption A 11 and also Assumption A 15 may be satisfied.

The intuition behind the two conditions in ((37) and (59)) is as follows. As we mentioned earlier, we introduced the design region $\mathcal{D}_n$ for each iteration $n$, in order to be able to control the Euclidean distances $\|y_n^i\|_2$ of the design points $x_n + y_n^i$ from the iterate $x_n$. Indeed we could, if we wanted, use only points within the design region $\mathcal{D}_n$ to determine $\hat{\nabla}_n f(x_n)$ and $\hat{\nabla}_n^2 f(x_n)$ for each $n \in \mathbb{N}$. However, as our optimization algorithm progresses, it is likely that some points in the set $\mathcal{X}_n$ (i.e., points at which sample averages have been evaluated in earlier iterations), will lie outside the design region $\mathcal{D}_n$ for iteration $n$. Since we have already evaluated sample averages at these points, we wish to include these points in the determination of $m_n$. However, in order for $\hat{\nabla}_n f(x_n)$ to better approximate $\nabla f(x_n)$, it is also true that the inner design points must provide progressively higher contributions to the determination $\hat{\nabla}_n f(x_n)$ and $\hat{\nabla}_n^2 f(x_n)$ as compared to the outer design points. Thus, the conditions (37) and (59), balance these two conflicting requirements by requiring that the outer design points, be assigned progressively smaller weights $\{w_n^{M_n^I+1}, \ldots, w_n^{M_n}\}$ as $n \rightarrow \infty$, when compared to the weights $\{w_n^1, \ldots, w_n^{M_n^I}\}$ assigned to inner design points. Essentially these
conditions ensure that as \( n \to \infty \), the outer design points are ignored and only the points within the design region are used to calculate \( \hat{\nabla}_n f(x_n) \) and \( \hat{\nabla}_n^2 f(x_n) \).

Consider the following method for setting the weights \( \{w_n^i : i = 1, \ldots, M_n\} \).

**Algorithm 10.** Let us suppose that the design points \( \{x_n + y_n^i : i = 1, \ldots, M_n\} \) are available to us. Let \( r \in \mathbb{R} \) be a positive constant.

**Step 1:** Define \( \pi_n \leftarrow \min_{i \in \{1, \ldots, M_n\}} \|y_n^i\|_2 \)

**Step 2:** Set for each \( i = 1, \ldots, M_n \),

\[
    w_n^i \leftarrow \begin{cases} 
        1 & \text{for } i \in \{1, \ldots, M_n\} \\
        \left( \frac{\pi_n}{\|y_n^i\|_2} \right) \min \{1, (\delta_n)^r\} & \text{for } i \in \{M_n + 1, \ldots, M_n\}
    \end{cases}
\]

(190)

Algorithm 10 warrants the following comments.

- In order for (190) to be well-defined, we must have \( \|y_n^i\|_2 > 0 \) for each \( i = M_n + 1, \ldots, M_n \). From our convention since \( \{x_n + y_n^i : i = M_n + 1, \ldots, M_n\} \) is the set of outer design points, we have \( \|y_n^i\|_2 \geq \delta_n \) for each \( i = M_n + 1, M_n \). Therefore, as long as \( \delta_n > 0 \), Algorithm 10 is well defined.

- It is easily seen that in the above example, \( w_n^i < 1 \) for each \( i = M_n + 1, \ldots, M_n \) and \( n \in \mathbb{N} \). Thus, the outer design points always get a lower weight as compared to the inner design points. Further, it is easy to see that if \( \delta_n \to 0 \) as \( n \to \infty \), then the weights given to the outer design points decrease to zero as \( n \to \infty \).

**Lemma 3.24.** Let the following assumptions hold.

1. The sequence \( \{\delta_n\}_{n \in \mathbb{N}} \) of design region radii satisfies \( \delta_n \to 0 \) as \( n \to \infty \).

2. There exists a constant \( M_{\max}^D < \infty \), such that \( 1 \leq M_n^I \leq M_n \leq M_{\max}^D \) for each \( n \in \mathbb{N} \).

Then, if the weights \( \{w_n^i : i = 1, \ldots, M_n\} \) are assigned using Algorithm 10 for each \( n \in \mathbb{N} \), (37) in Assumption A 7 is satisfied for any \( r > 0 \).

Further, if there exists \( K_y < \infty \) such that \( \|y_n\|_2 < K_y \) for all \( i = 1, \ldots, M_n \) and \( n \in \mathbb{N} \), then (59) in Assumption A 11 is satisfied for any \( r > 2 \).

**Proof.** First, let us consider (37). It is not hard to see that

\[
    \frac{\|\hat{Y}_n^T W_n^O \hat{\nabla}_n^O\|_2^2}{\|\hat{Y}_n^T W_n^I \hat{\nabla}_n^I\|_2^2} = \frac{\|Y_n^O T W_n^O Y_n^O\|_2^2}{\|Y_n^I T W_n^I Y_n^I\|_2^2} \leq l \left( \frac{\text{trace}(Y_n^O T W_n^O Y_n^O)}{\text{trace}(Y_n^I T W_n^I Y_n^I)} \right) = l \left( \frac{\sum_{i=M_n+1}^{M_n} w_n^i \|y_n^i\|_2^2}{\sum_{i=1}^{M_n} w_n^i \|y_n^i\|_2^2} \right)
\]

93
Substituting the weights mentioned in (190) into the above and manipulating the resulting expressions, we get
\[ \left\| (\hat{Y}^O_n)^T W^O_n \hat{Y}^O_n \right\|_2 \leq l \left( \frac{\min \{1, (\delta_n)^r \} \pi^2_n (M_n - M^I_n)}{\sum_{i=1}^{M^I_n} \| y^i_n \|^2_2} \right) \leq l \left( \frac{\min \{1, (\delta_n)^r \} (M_n - M^I_n)}{M^I_n} \right) \]

We know that \( M^D_{\text{max}} \geq M_n \geq M^I_n \geq 1 \) for each \( n \in \mathbb{N} \). Therefore, \( (M_n - M^I_n)/M^I_n \leq 2M^D_{\text{max}} \). Further, since \( \delta_n \to 0 \) as \( n \to \infty \) and \( r > 0 \), we finally get
\[ \lim_{n \to \infty} \left\| (\hat{Y}^O_n)^T W^O_n \hat{Y}^O_n \right\|_2 \leq 2l \left( M^D_{\text{max}} \lim_{n \to \infty} \min \{1, (\delta_n)^r \} \right) = 0 \]

Thus, (37) in Assumption 7 is satisfied.

Next, assuming in addition, that \( \| y^i_n \|_2 < K_y \) for all \( i = 1, \ldots, M_n \) and \( n \in \mathbb{N} \), consider (59). It is easily seen that
\[ \left\| (\bar{Z}^O_n)^T W^O_n \bar{Z}^O_n \right\|_2 \leq p \left( \frac{\sum_{i=M^I_n+1}^{M_n} w^i_n \| z^i_n \|^2_2}{\sum_{i=1}^{M^I_n} w^i_n \| z^i_n \|^2_2} \right) \]

Now, substituting the weight distribution from (190) into the numerator of the above expression and using the definition of \( z^i_n \), we get
\[ \sum_{i=M^I_n+1}^{M_n} w^i_n \| z^i_n \|^2_2 = \sum_{i=M^I_n+1}^{M_n} \min \{1, (\delta_n)^r \} \left( \frac{\pi^2_n}{\| y^i_n \|^2_2} \right) \| z^i_n \|^2_2 \leq \sum_{i=M^I_n+1}^{M_n} (\delta_n)^{r-2} \pi^2_n \left( \frac{\| y^i_n \|^2_2 + \| y^i_n \|^4_2}{2\delta_n^2} \right) \]
\[ \leq \sum_{i=M^I_n+1}^{M_n} (\delta_n)^{(r-2)} \pi^2_n \left( 1 + \frac{\| y^i_n \|^2_2}{2\delta_n^2} \right) \]

Also, we have
\[ \sum_{i=1}^{M^I_n} w^i_n \| z^i_n \|^2_2 = \sum_{i=1}^{M^I_n} \left( \left( \frac{\| y^i_n \|^2_2}{\delta_n^2} + \frac{\| y^i_n \|^4_2}{2\delta_n^4} \right) \right) \geq \sum_{i=1}^{M^I_n} \left( \frac{\| y^i_n \|^2_2}{\delta_n^2} \right) \]

Thus, combining (191) and (192) and noting that \( \| y^i_n \|_2 < K_y \) for all \( i = 1, \ldots, M_n \) and \( n \in \mathbb{N} \), we have
\[ \sum_{i=M^I_n+1}^{M_n} w^i_n \| z^i_n \|^2_2 \leq \sum_{i=M^I_n+1}^{M_n} \left( \delta_n^{(r-2)} \pi^2_n \right) \left( 1 + \frac{\| y^i_n \|^2_2}{2\delta_n^2} \right) \leq (\delta_n)^{(r-2)} \frac{(\delta_n^2 + K_y^2)}{2} (M_n - M^I_n) \]
\[ \sum_{i=1}^{M^I_n} \frac{\| y^i_n \|^2_2}{\delta_n^2} \]

As before, since \( (M_n - M^I_n)/M^I_n \leq 2M^D_{\text{max}} \), \( \delta_n \to 0 \) as \( n \to 0 \) and \( r > 2 \), we get
\[ \lim_{n \to \infty} \sum_{i=M^I_n+1}^{M_n} w^i_n \| z^i_n \|^2_2 \leq 2M^D_{\text{max}} \lim_{n \to \infty} (\delta_n)^{(r-2)} \left( \delta_n^2 + \frac{K_y^2}{2} \right) = 0 \]

Thus, we finally get that
\[ \lim_{n \to \infty} \left\| (\bar{Z}^O_n)^T W^O_n \bar{Z}^O_n \right\|_2 \leq p \lim_{n \to \infty} \left( \frac{\sum_{i=M^I_n+1}^{M_n} w^i_n \| z^i_n \|^2_2}{\sum_{i=1}^{M^I_n} w^i_n \| z^i_n \|^2_2} \right) = 0 \]
Note the following regarding the assumptions of Lemma 3.24.

1. Suppose we pick our design points \( \{x_n + y^i_n : i = 1, \ldots, M_n\} \) for each \( n \in \mathbb{N} \) using either Algorithm 4 or 7. Then for each \( n \in \mathbb{N} \), we note that before either of these algorithms are executed, we know that the number of entries in \( \mathcal{X}_n \) is at most \( M_{\text{max}} \). In either Algorithm 4 we add at most \( l \) points into \( \mathcal{X}_n \), and in Algorithm 7 we add at most \( p \) new points in to \( \mathcal{X}_n \). Therefore, in either case, it is easily seen that \( l \leq M_n^I \leq M_n \leq M_{\text{max}} + p \).

2. Further, it is easily seen that if there exists a compact set \( \mathcal{C} \subset \mathcal{E} \) such that \( x_n + y^i_n \in \mathcal{C} \) for each \( n \in \mathbb{N} \) and \( i = 1, \ldots, M_n \), then there exists \( K_y < \infty \) such that \( \|y^i_n\|_2 < K_y \) for each \( n \in \mathbb{N} \) and \( i = 1, \ldots, M_n \).

Finally it is also clear that if the weights \( \{w^i_n : i = 1, \ldots, M_n^I\} \) corresponding to the inner design points, are picked as in Algorithm 10, then

\[
\min\{w^i_n : i = 1, \ldots, M_n^I\} = \max\{w^i_n : i = 1, \ldots, M_n^I\} = 1
\]

Therefore, in this case Assumption A 15 used in Section 3.2.1 is satisfied with \( K_{w^I} = 1 \).