International Journal of Advanced Research in Engineering and Applied Sciences
ISSN: 2278-6252
Impact Factor: 7.358

NON LINEAR FUNCTION'S OPTIMIZATION WITH GLOBAL CONVERGENCE ON PROBABILISTIC APPROACH BY DIFFERENT LATEST TECHNIQUES

Guash Haile Taddele, College of Natural & Computation Sciences, Department Head, Department of Mathematics, DBU

Abstract: *From the Revolution of Mathematics by Newton's Calculus most of mathematicians, Scholars, researchers try to find optimization of Linear and Non Linear Problems (NLP). Up to 19th century most of researchers developed many techniques for solving linear types only. After Hilbert and Banach implementationsespecially in Functional Analysis and Theory of Approximations, scholars focused their mission in NLP. Kuhn – Tucker, wolf's identified and applied their own ideas to solve NLP. In this paper we try to solve NLP by Trust – Region methods, Rate Analysis of Unconstrained Methods, Linear Search Concepts and Direct Search Approach by Conjugate Gradient Concept with Globally convergence*

Keywords: *Minimum Norm Gradient, Generic Probabilistic, Noise Scales, Line Searches Strategy, Hyper Parameters*

INTRODUCTION AND MOTIVATION

In this article we will discuss about is the analysis of a numerical scheme that utilized models to minimize determined functions. In particular, our aim comes from algorithms for minimization of **Block – Box** functions where values are computed, for example via simulations. For such problems, function evaluations are costly and derivatives are typically unavailable and cannot be approximated. Such is the setting of **Derivative – Free Optimization** (DFO) **[8](Category – 1)**

Secondly, the convergence properties of several conjugate gradient methods for nonlinear optimization. **(Category – 2)**: We consider only the case where the methods are implemented without regular restarts, and ask under what conditions they are globally convergent for general smooth nonlinear functions. The analysis will allow us to highlight differences among various conjugate gradient methods, and will suggest new implementations. In this case our problem is minimize a function of variables,

$$\min f(x) \longrightarrow \quad (2.1)$$

Where f, it is a smooth, and its gradient g, it is available and consider the iterations of the form

$$d_k = \begin{cases} -g_k, & for k = 1 \\ -g_k + \beta_k d_{k-1}, & for k \geq 2 \end{cases} \longrightarrow (2.2)$$

With $x_{k+1} = x_k + \alpha_k x_{k-1}$ $\longrightarrow$ (2.3)

Where β_k, it is a scalar and α_k, it is a step length 0btained by means of one – dimensional search. We call this iteration "A Conjugate Gradient Method", if β_k, it is such that (2.2) and (2.3) reduces to the linear conjugate gradient method in the case when f, it is a strictly convex quadratic and α_k, it is the exact one – dimensional minimize. Some of the results of this article, however, also apply to methods of form (2.2) and (2.3) that o not reduce to the linear conjugate gradient method. The best known formula for β_k, they are called the Fletcher – Reeves (FR) [13], Polak – Ribiere (PR) [24] and Hestense – Stiefel (HS) [14, 17] formula are they given by

$$\beta_k^{FR} = \|g_k\|^2 / \|g_{k-1}\|^2 \longrightarrow (2.4)$$

$$\beta_k^{PR} = \langle g_k, g_k - g_{k-1} \rangle / \|g_{k-1}\|^2 \longrightarrow (2.5)$$

$$\beta_k^{HS} = \langle g_k, g_k - g_{k-1} \rangle / \langle d_{k-1}, g_k - g_{k-1} \rangle \longrightarrow (2.6)$$

Here $\langle . \rangle$, it is the scalar product used to compute the gradient and $\|.\|$, denotes the associated norm. Note that the numerical performance of FR method is somewhat erratic, it is sometimes as efficient as PR and HS methods, but it is often slower. Powell (1977) [25] showed that, under some circumstance, the FR method with exact lines searches will produce very small displacements and will normally not recover unless a restart along the gradient direction is performed. This drawback will be cured by Zoutendijk (1970)[38], FR method with exact line searches is globally convergent on the gradient functions. Al – Baali (1985)[1] extended this result to inexact line searches. In this paper we will consider various choices of β_k, and various line search strategies that result in globally convergent methods. Our assumptions in (2.4) it is $|\beta_k| \leq \beta_k^{FR}$, it describes the modified the PR formula and in (2.5) we consider only non – negative values for β_k, and these are some sense, related to PR method. In particular we show that a suggestion of Powell (1985) [27], to set $\beta_k = \max\{\beta_k^{PR}, 0\}$, results in global convergence, even for inexact line searches. Further remarks on the convergence results are made in (2.5), and the results of some numerical experiments are presented in (2.6). Finally we note that this article does not study the rate of convergence of conjugate gradient methods.

Trust – Region Framework – (TR):[4]

TRM introduced and analyzed at each iteration one solve a TR sub problem, i.e. one minimizes the model within a TR ball. Note that one does know whether the model is accurate or not. If the TR step yields a good decrease in the objective function relatively to the decrease in the model and the TR radius is sufficiently small relatively to the size of the model gradient, then the step is taken and the TR radius is possibly increased. Otherwise the step is rejected and the TR radius is decreased. We show that such a method always drive the TR radius to zero. Based on the property we show that, provided the (First Order) accuracy of the model occurs with probability no smaller that$\frac{1}{2}$, conditioned to the prior iteration history, then the gradient of the objective function converges to zero with probability 1. Our proof techniques relies on building random process from the random events defined by the models being or not being accurate, and then making use of their sub martingale – like properties. We can extend this model of sufficient second order accuracy occur with probability no smaller than 0.5. We show that a subsequenceof iterates drive a measure of second order stationary to zero with probability 1. However, to demonstrate the limit – type convergence to a second order stationary point we need additional assumptions on the model. We discussed only First order scheme.

Methods of Derivate – Free Optimization (DFO):[2]

Consider the unconstrained optimization problem

$\min_{x\in\mathbb{R}^n} f(x)\ where\ f(x)$, it is the first and second derivatives of the objective function and assumed to exist and be Lipchitz Conditions, however as it is considered in DFO , explicit evaluation of these derivatives is assumed to be impossible. Derivative – Free methods relay on sampling the objective function ateither one or more points,at eachof iteration and some sample to explore directions, other to build models.

Directional Search Methods:

Among the method of directional type to minimization without derivatives are the direct – search methods with were developed using a single positive spanning set or a finite number of them (See surveys [12] and [8], Chapter – 8). On the other hand, randomized stochastic methods recently became a popular alternative to direct – search methods. These methods are also directional, but instead of using directions from a positive spanning set, they select a search direction randomly. This can allow faster convergent because directions of significant descent may be occasionally observed, which might not be the case when insisting on using directions from a fixed positive spanning set (and the use of a randomly rotated positive

spanning set may require polling all its directions to find such a direction of significant descent). The random search approach introduced in [19] samples points from a Gaussian distribution.

Model – based on TRM:

Model – based DFO methods developed by Powell [26, 27, 28, and 29] and by Conn, Scheinberg, and Toint [5, 6] introduced a class of TRM that relied on interpolation or regression based quadratic approximations of the objective function instead of the usual Taylor series quadratic approximation. The regression – based method was later successfully used in[3] based on [7]. In all cases the models are built based on sample points in reasonable proximity to the current best iterate. The computational study of Moor and Wild [23] has shown that these methods are typically significantly superior in practical performance to the other existing approaches due to the use of models that effectively captured the local curvature of the objective function. While the model quality is undoubtedly essential for the performance of these methods, guaranteeing sufficient quality on specific iterations is quite expensive computationally. Randomized models, on the other hand, can offer a suitable alternative by providing a good quality approximation with high probability.

An illustration of directional and model – based methods:

Consider the well known Rosenbrock function for our computational illustration

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_2)^2$$

The function is known to be difficult for first order or zero order methods and well suited for second order methods. Nevertheless some first/zero order methods performs reasonably, while others performs poorly. We compared the following four methods.

1. A simple variant of direct search, the coordinate or compass search method (CS), it uses the positive basis$[I \ -I], wher\ I$, it is the identity matrix (CS)
2. A direct –search method using positive basis$[Q - Q]\ where\ Q$, it is an orthogonal matrix obtained by randomly generating the first column (DSR)
3. A random search (RS) with step size inversely proportional to the iterations count and
4. A basic model – based trust – region method with quadratic model (TRQ)

The outcome of the algorithms is summarized as follows:

Method	No. of function Evaluation	Final Function Value
CS	11307	$1.0e-6$
DSR	5756	$1.0e-8$
RS	3724	$1.0e-8$
TRQ	62	$1.0e-14$

That is, in particular random research are more successful at finding good directions for descent, while the coordinate search is slow due to the fixed choice of the search directions. It is also clear, from the performance of the second order trust – region method on this problem, that using accurate models can substantially improve efficiency. It is natural, thus, to consider the effects of randomization in model – based methods. In particular we consider methods that use models built from randomly sampled points in hopes of obtaining better models.

First order trust – region method based on probabilistic models:

Consider the classical trust – region method setting and notation (see [8]), at iteration k, f, it is approximated by a model$m_k$within the ball $B(x_k, \delta_k)\ centered\ at\ x_k\ and\ radius\ \delta_k$. Then the model is minimized or approximately minimized in the ball to possibly obtainx_{k+1}. In this article we will introduce and analyze a trust – region algorithm based on probabilistic models i.e. modelsm_k, they are built in a random fashion. We will discuss this models and state what will be assumed from them.

The Probabilistically fully linear models:

Consider a quadratic model written in the form$m_k(x_k + s) = m_k(x_k) + s^T g_k + \frac{1}{2} s^T H_k s$

Where$g_k = \nabla m_k(x_k)\ \ and\ H_k = \nabla^2 m_k(x_k)$: Our analysis is not, however, dependent on the models being quadratic and introduces a measure of (linear or first order) accuracy of the model m_k

Definition [7, 8, 9]:

A functionm_k, it is called$(\kappa_{eg}, \kappa_{ef})$,*fully – linear model*$f\ on\ B(x_k, \delta_k)\ where$ $\kappa_{eg}\ is\ error in gradient, \kappa_{ef}\ is\ error in function value$, if $\forall\ s \in B(0, \delta_k)$, and then we have

$$\|\nabla f(x_k + s) - \nabla m_k(x_k + s)\| \leq \kappa_{eg}\, \delta_k\, and \|f(x_k + s) - m(x_k + s)\| \leq \kappa_{ef}\, \delta_k^2$$

Consider the random models $M_k\ \ and\ take\ m_k = M_k(\omega_k)$, for their realizations. The randomness of the models will imply the randomness of pointsx_k and the trust region radiiδ_k. Thus, in the sequel, these random quantities denoted by$X_k\ \ and\ \nabla_k$respectively while$x_k = X_k(\omega_k)\ \ and\ \delta_k = \nabla_k(\omega_k)$, denote their realization.

Definition [8]

A sequence of random models$\{X_k\}$, it is$(p) - probabilistically(\kappa_{eg}, \kappa_{ef})$- *fully linear* for a corresponding sequence$\{B(X_k, \nabla_k)\}$, if the events

$S_k = \{M_k\ is a (\kappa_{eg}, \kappa_{ef}) - fully linear arm model of f\ on B(X_k, \nabla_k)\}$,satisfies the following sib martingale – like condition $P(S_k | F_{k-1}^M) \geq p where F_{k-1}^M = \sigma(M_0, \ldots, M_{k-1})$, it is

the $\sigma-$algebra generated by $(M_0, \dots, M_{k-1})$ furthermore, if $p \geq \frac{1}{2}$, and then we say that the random models are probabilistically $(\kappa_{eg}, \kappa_{ef})$- fully linear

Assumption – 1:

$\forall\ k$, and for all realizations $m_k of M_k and X_k of \nabla_k$, we can compute a steps s_k such that

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{\kappa_{fcd}}{2} \|g_k\| \min\left\{\frac{\|g_k\|}{\|H_k\|}, \delta_k\right\} \longrightarrow \quad (1.1)$$

For some constant κ_{fcd} (Fraction of Cauchy Decrease)$\in (0,1]$, in this case that s_k, it has achieved a fraction of Cauchy decrease, the Cauchy step itself, which is the minimize of the quadratic model within the trust region along the negative model gradient $-g_k$, trivially satisfies this property with $\kappa_{fcd} = 1$

Assumption – 2:

There exists a constant $\kappa_{bhm} > 0$ (Bound on the Hessian of the Models), such that $\forall\ k$, the Hessians H_k of all realizations $m_k\ of\ M_k$ satisfy

$$\|H_k\| \leq \kappa_{fcd} \longrightarrow \quad (1.2)$$

This assumption is introduced for convenience. What it is possible to show our results without this assumption, it is not restrictive in the case of fully linear models. In particular, one can construct fully linear models with arbitrarily small $\|H_k\|$, using interpolation techniques. In the case of models that have large Hessian norms, because they are not fully linear, we can set the Hessian to some other matrix of a smaller norm.

Algorithm – 1 and Basic Properties:

Consider the simple trust – region algorithm, fix the positive parameters $\eta_1, \eta_2, \gamma, \delta_{max}\ with \gamma > 1 > \eta_1$. Select initial $k = 0, \delta_0 \leq \delta_{max}\ and x_0$. At iteration k approximate $f in B(x_k, \delta_k)$ by m_k, and then approximately minimize $m_k in B(x_k, \delta_k)$, computing s_k so that it satisfies a function of Cauchy decreases (3) and let

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m(x_k) - m(x_k + s_k)} \longrightarrow \quad (1.3)$$

If $\rho_k \geq \eta_1$, then set $x_{k+1} = x_k + s_k and \delta_{k+1} = \begin{cases} \gamma^{-1}\delta_k, & if \|g_k\| < \eta_2 \delta_k \\ min\{\gamma\delta_k, \}\delta_{max}, & if \|g_k\| \geq \eta_2 \delta_k \end{cases}$

Otherwise, set $x_{k+1} = x_k and \delta_{k+1} = \gamma^{-1}\delta_k$. Increasing $k\ \ by\ 1$ and repeat the iteration

Explanation:

This is a basic TR algorithm, with one specific modification the TR radius is always increased if sufficient function reduction is achieved, i.e. the step is successful, and the TR radius is small compared to the norm of the model gradient. The logic behind this update follows from the line – search (second concept) type intuition, where the step size is typically

proportional to the norm of the model gradient, hence the TR should be of comparable size also. Later we will show how the algorithm can be modified to allow for the TR radius remain unchanged in some iterations.Each realization of the algorithm defines a sequence of realizations for the corresponding random variables, in particular:

$$m_k = M_k(\omega_k), x_k = X_k(\omega_k) \ and \ \delta_k = \nabla_k(\omega_k)$$

For the purpose of proving convergence of the algorithm to first order critical points, we assume that the function f and its gradient are Lipchitz continuous in region s considered by the algorithm realization. To define this region we follow the process [8]. Suppose thatx_0 (Initial iterate) is given. Then all the subsequent iteratesbelong to the level set $L(x_0) = \{x \in \mathbb{R}^n | f(x) \leq f(x_0)\}$

However, the failed iterates may lie outside this set. In the setting considered in this paper, all potential iterates are restricted to the region

$L_{enl}(x_0) = L(x_0) \cup_{x \in L(x_0)} B(x, \delta_{max}) = \cup_{x \in L(x_0)}(x, \delta_{max}) where \delta_{max}$, it is the upper bound on the size of TR, as imposed by the algorithm – 1.

Assumption – 3:

Suppose $x_0 \ and \ \delta_{max}$, they are given: Assume that f, it is continuously differentiable in an open set containing the set$L_{enl}(x_0)$ and that∇f, it is Lipchitz continuous on$L_{enl}(x_0)$ with constant κ_{Lg}(the Lipchitz constant of the gradient function) and also f, it is bounded below on $L(x_0)$ and we need the **lemma – 1 :** "For every realization of our algorithm $\lim_{k \to \infty} \delta_k = 0$" [15] and

Lemma - 2: "If m_k , it is $(\kappa_{eg}, \kappa_{ef})$,fully – linear model f on $B(x_k, \delta_k)$ and $\delta_k \leq \min\left\{\frac{\|g_k\|}{\kappa_{bhm}}, \frac{\kappa_{fcd}(1-\eta_1)\|g_k\|}{4\kappa_{ef}}\right\}$, and then at the kth iteration $\rho_k \geq \eta_1$"[8]

Convergence of the first order TR method based on probabilistic models:

Assume that the model used in the algorithm are probabilistically fully linear, and show our first order convergence results, so that we state an auxiliary results from the Martingale literature that is

Theorem – 1: [10]

LetG_k, it is a sub martingale, i.e. a sequence of random variables which, $\forall \ k$, they are integrable $(\mathbb{E}(|G_k|) < \infty) and \ \mathbb{E}[G_k | F_{k-1}^G] \geq G_{k-1} where F_{k-1}^G = \sigma(G_0, \dots, G_{k-1})$, it is theσ −algebra, generated by$G_0, \dots, G_{k-1} and \ \mathbb{E}[G_k | F_{k-1}^G]$denotes the conditional expectation ofG_k, given the past history of eventsF_{k-1}^G, assume that$|G_k - G_{k-1} \leq M < \infty, \forall \ k|$. Consider the random events

$$C = \left\{\lim_{k\to\infty} G_k exists and finte\right\} and D = \left\{\lim_{k\to\infty} G_k = \infty\right\} then P(C \cup D) = 1$$

The Limit – Type Convergence:

In TR methods, we show first that a subsequence of iterates drive the gradient of the objective function to zero by theorem – 2:

Theorem – 2:[9, 22]

Suppose that the model sequence $\{M_k\}$, it is $(\kappa_{eg}, \kappa_{ef})$,fully – linear for some positive constant $\kappa_{eg}\ and\ \kappa_{ef}$. Let$\{X_k\}$, it is a sequence of random iterates generated by algorithm – 1 and then almost surely $\lim_{k\to\infty} \inf\|\nabla f(X_k)\| = 0$

This results achieved by following two lemmas the proofs are available in [9, 22]

Lemma – 3: [22]

Let$\{Z_k\}_{k\in\mathbb{N}}$, it is a sequence of non – negative uniformly bounded random variables and$\{B_k\}$, it is a sequence of Bernoulli random variables (taking values $1\ and\ -1$) such that

$$P\left(B_k = 1|\sigma(B_1, \dots, B_{k-1}), \sigma(Z_1, \dots, Z_k)\right) \geq \frac{1}{2}$$

Let $\mathcal{P}$, it is the set of natural numbers$k$ such that$B_k = 1\ \ and\ \mathcal{N} = \mathbb{N}/\mathcal{P}$, note that $\mathcal{P}\ \ and\ \mathcal{N}$ they are random sequences then$Prob(\{\sum_{i\in\mathcal{P}} Z_i < \infty\} \cap \{\sum_{i\in\mathcal{N}} Z_i = \infty\}) = 0$

Lemma – 4:[22]

Let$\{X_k\}\ \ and\ \{\Delta_k\}$, they are sequence of random iterates and random TR radii generated by algorithm – 1. Fix $\varepsilon > 0$and define the sequence$\{K_i\}$ consisting of the natural numbers k for which$\|\ \ \|\nabla f(X_k) > \varepsilon$ (note that K_i, it is sequence of random variables). Then$\sum_{k\in\{K_i\}} \Delta_k <$ ∞, almost surely

Theorem for limit – type result – 3:

Suppose that the model sequence $\{M_k\}$, it is $(\kappa_{eg}, \kappa_{ef})$,fully – linear for some positive constant $\kappa_{eg}\ and\ \kappa_{ef}$. Let$\{X_k\}$, it is a sequence of random iterates generated by algorithm – 1 and then almost surely $\lim_{k\to\infty} \|\nabla f(X_k)\| = 0$ (See the proof) [22]

Modified TR Schemes:

The TR radius update of $(\boldsymbol{Algorithm - 1})$may be too restrictive as it only allow for this radius to be increased or decreased. In practice typically two separate thresholds are used, one for the increase of the TR radius and another for its decrease. In the remaining cases the TR radius remains unchanged. Hence, we propose an algorithm similar to Algorithm – 1, but slightly more appealing in practice.

Algorithm – 2:

Fix the positive parameters$\eta_1, \eta_2, \eta_3, \gamma, \delta_{max}\, with \gamma > 1 > \eta_1 and \eta_2 \leq \eta_3$. Select initial$k = 0, \delta_0 \leq \delta_{max}\, and x_0$. At iteration$k$approximate$f in B(x_k, \delta_k) by m_k$, and then approximately minimize$m_k in B(x_k, \delta_k)$, computings_k so that it satisfies a function of Cauchy decreases (3) and let$\rho_k$it is defined as in (5). If$\rho_k \geq \eta_1$ and then set$x_{k+1} = x_k + s_k and$

$$\delta_{k+1} = f(x) = \begin{cases} \gamma^{-1}\delta_k, & if \|g_k\| < \eta_3 \delta_k \\ \delta_k, & if \eta_3 \delta_k \leq \|g_k\| < \eta_2 \delta_k \\ min\{\gamma\delta_k, \delta_{max}\}, & if \eta_2 \delta_k \leq \|g_k\| \end{cases}$$

Otherwise set$x_{k+1} = x_k \; and \; \delta_{k+1} = \gamma^{-1}\delta_k$. Increase k by one and repeat the iteration

It is straightforward to adapt the proofs of lemma – 1 and theorem – 2 and 3 to show the convergence for this new algorithm – 2. Additionally, one can consider two different thresholds $0 < \eta_0 < 1$ for decrease of the TR radius, and$\eta_1 > \eta_0$for the increase of the TR radius.

The convergence properties of several conjugate gradient methods for nonlinear optimizations:

Background of studies or preliminaries:

Some important Global Convergence for Conjugate Gradient Methods (GCCGM) have been given by Polak and Ribiere (1969) [24] method (PR), Zoutendijk (1970)[38], Powell (1984) [26] and Al – Baali |(1985) [1]. In this article we will see that underlying approach used for these analyses is essentially the same, and we will describe it in detail, since it is also the basis for the result presented in this paper.

Definition:

Take starting pointx_1, and define$s_k = x_{k+1} - x_k and y_k = g_{k+1} - g_k$ ifd_k, it is a **Descent Direction** (DD) if$\langle g_k, d_k \rangle < 0$, and also the angle$\theta_k$, it is between $-g_k and d_k$, and then we have

$$\cos\theta_k = -\langle g_k, d_k \rangle / \|d_k\| \longrightarrow \quad (2.6)$$

The Fletcher – Reeves (FR), Polak – Ribiere (PR) and Hestenes – Stiefel (HS) methods have been discussed here.

Assumption – 2.1 for (Category 2):

1. The level set $\mathcal{L}(x) = \{x | f(x) \leq f(x_1)\}$ it is bounded
2. In some neighborhood $\mathcal{N}\, of\, \mathcal{L}$, the objective function f, it is continuous and differentiable, and its gradient is Lipchitz conditions i.e. there exists a constant $L > 0$ such that

$$\|g(x) - g(\tilde{x})\| \leq L\|x - \tilde{x}\|, \forall\, x, \tilde{x} \in \mathcal{N} \longrightarrow (2.7)$$

From these assumptions, there is a constant$\tilde{\gamma}$, such that$\|g(x)\| \leq \tilde{\gamma}, \forall\, x \in \mathcal{L} \longrightarrow (2.8)$

In line search, Wolf (1969) [36] consists in accepting a step length$\alpha_k > 0$, satisfies the two conditions

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \sigma_1 \alpha_k \langle g_k, d_k \rangle \longrightarrow (2.9)$$

$$\langle g(x_k + \alpha_k d_k), d_k \rangle \geq \sigma_2 \langle g_k, d_k \rangle \longrightarrow (2.10)$$

Where $0 < \sigma_1 < \sigma_2 < 1$, and we define the strategy: A step length$\alpha_k > 0$, it is accepted if

$$f(x_k + \alpha_k d_k) \leq f(x_k + \hat{\alpha}_k d_k) \longrightarrow (2.11)$$

($where\ \hat{\alpha}_k$, it is the smallest positive stationary point of the function$\xi_k(\alpha) = f(x_k + \alpha_k d_k)$

By assumption – 2.1$\hat{\alpha}_k$, it exists and also both first local minimize as well as the global minimize of f, along the search direction satisfy (2.11)

Theorem – 2.1(Zoutendijk Condition)[38]

Suppose that assumption 2.1 holds and consider any iteration of the form (2.3), whered_k, it is the descent direction andα_k satisfies one of the following line search conditions

1. The Wolf conditions (2.9) and (2.10) or
2. The ideal line search condition (2.11) and then

$$\sum_{k\geq 1} cos^2\theta_k \|g_k\|^2 < \infty \longrightarrow (2.12)$$

From this condition we describe the basic idea used for convergence analysis, the first result by PR, they assume exact line searches. The term exact line search can be ambiguous; it implies that one dimensional minimizer is bound that is the orthogonality condition

$$\langle g_k, d_{k-1} \rangle = 0 \longrightarrow (2.13)$$

The whole article we will indicate in detail the conditions required of the line search, suppose thatd_{k-1} satisfies Zoutendijk's condition and (2.13) we have

$$\cos\theta_k = \|g_k\|/\|d_k\| \longrightarrow (2.14)$$

$\Longrightarrow d_k$, it is a descent direction and we substitute (2.14) in (2.7) we get

$$\sum_{k\geq 1} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty \longrightarrow (2.15)$$

$\Longrightarrow \{\|d_k\|/\|g_k\|\}$, it is bounded$\Longrightarrow \{\cos\theta_k\}$, it is bounded away from zero then by (2.15) we get

$$\lim_{k\to\infty} g_k = 0 \longrightarrow (2.16)$$

This is done by Polak and Ribiere (1969) [24]and assumes that f , it is strongly convex, i.e.

$\langle g(x) - g(\tilde{x}), x - \tilde{x}\rangle \geq c\|x - \tilde{x}\|^2 \; for\; some\; c > 0\; and\; \forall\; x, \tilde{x} \in \mathcal{L}$, and for general functions, however it is impossible to bound$\{\|d_k\|/\|g_k\|\}$ a priori and only a weaker result in (2.16) can be obtained by

$$\lim_{k\to\infty} \inf\|g_k\| = 0 \longrightarrow \quad (2.17)$$

To obtain this result one proceeds by contradiction, suppose that (2.16) does not hold, the gradients remain bounded away from zero, there exists $\gamma > 0$ such that

$$\|g_k\| \geq \gamma, \forall\; k \geq 1 \longrightarrow \quad (2.18)$$

$$\Longrightarrow \sum_{k\geq 1} \frac{1}{\|d_k\|^2} < \infty \longrightarrow \quad (2.19)$$

Conclusion: The iteration fails only if $\|d_k\| \longrightarrow \infty$, sufficiently rapidly, i.e. (2.18) holds then$\|d_k\|^2$, it can grow at most linearly i.e.$\|d_k\|^2 \leq ck, for\; some\; constant\; c$

This contradicts to (2.19), proving ((2.17). That is the analysis for inexact line search that satisfies Zoutendijk's condition can proceed along the same line if that iteration satisfies

$$\cos\theta_k \geq c\,\|g_k\|/\|d_k\|\, forsomeconstantc \longrightarrow \quad (2.20)$$

Then, this relation can be used instead of (2.14) to give (2.15), and the rest of analysis is as in the case of exact line search. Al – Baali (1985) [1], shows that FR methods give (2.20), if the step length satisfies the strong Wolf condition:[37]

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \sigma_1\alpha_k\langle g_k, d_k\rangle \longrightarrow \quad (2.21)$$

$$\langle g(x_k + \alpha_k d_k), d_k\rangle \leq -\sigma_2\langle g_k, d_k\rangle \longrightarrow \quad (2.22)$$

Where $0 < \sigma_1 < \sigma_2 < 1$, in fact it is necessary to require that $\sigma_2 < \frac{1}{2}$, for the result holds for FR method.

Al – Baali's result explained (2.20) which is (2.6) and equivalent to

$$\langle g_k, d_k\rangle \leq -c\|g_k\|^2 \longrightarrow \quad (2.23)$$

And also FR method using the strong Wolf conditions with $\sigma_2 < \frac{1}{2}$always generates descent directions. In this article we use the approach described above to establish the global convergence of various algorithms with inexact line searches. We will repeatedly encountered (2.23), it appears to be a natural way of guaranteeing descent for conjugate gradient methods and we called (2.23), *the sufficient condition*. We also show that any method of the form (2.2), (2.3), they globally convergent ifβ_k satisfies$|\beta_k| \leq \beta_k^{FR}$

This result suggests a new implementation of the PR method that preserves its efficiency and assures its convergence. We also study methods with$\beta_k \geq 0$, it is in some sense, related to PR method. A particular case is the following adaptation of the PR method, it consists in restricting $\beta_k > 0$ and let

$\beta_k = \max\{\beta_k^{FR}, 0\}$ ⟶ (2.24)

This motivation for this strategy arises from Powell's analysis of the PR method, he assumes that the line search always finds the first stationary point, and shows that there is a twice continuously differentiable function and a starting point such that the sequence of gradients generated by the PR method stays bounded away from zero. Since Powell example requires that some consecutive search directions become almost contrary, and since this can only be achieved in the case exact line searches when $\beta_k < 0$, Powell suggests modifying the PR method as in (2.24).

We show that this choice of $\beta_k$it does indeed result in global convergence, both for exact and inexact line search. Moreover we show that the analysis also applies to a family of methods with $\beta_k \geq 0$ that share a common property with the PR method and called**Property (*)**

Iterations constrained by the FR method:

We will see that it is possible to obtain global convergence if the parameter β_k, bounded in magnitude.

Consider the method of the form (2.2), (2.3) where β_k, it is any scalar such that

$|\beta_k| \leq \beta_k^{FR}, \forall\, k > 2$ ⟶ (2.25)

And where the step length satisfies the strong Wolf conditions (2.21), (2.22) with $\sigma_2 < \frac{1}{2}$, in this case theor3m 2.1 holds and also satisfies the Wolf Conditions[33](2.9) and (2.10), we need the following two results for our research

Lemma – 2.1:Touati – Ahamed and Stroey (1990) [34]:

Suppose that the assumption 2.1 holds and consider any method of the form (2.2), (2.3) whereβ_k, satisfies (2.25), and where step length satisfies the Wolf conditions (2.21), (2.22) with $\sigma_2 < \frac{1}{2}$, and then the method generates descent directions d_k satisfying

$-\sum_{j=0}^{k-1} \sigma_2^j \leq \frac{\langle g_k, d_k \rangle}{\|g_k\|^2} \leq -2 + \sum_{j=0}^{k-1} \sigma_2^j \,, for k = 1, \ldots,$ ⟶ (2.26)

This lemma 2.1, achieves three objectives:

1. It shows that all search directions are descent directions and the upper bound in(3.26) shows that the sufficient condition(2.23) holds
2. The bounds on$\langle g_k, d_k \rangle$ impose a limit on how fast$\|d_k\|$ it can grow when the gradient are not small, as we will see in Theorem 2.2
3. From (2.1) and (2.26) there exist$c_1, c_2 > 0$, such that$c_1 \frac{\|g_k\|}{\|d_k\|} \leq \cos\theta_k \leq c_2 \frac{\|g_k\|}{\|d_k\|}$ ⟶ (2.27)

Therefore, for the FR method with$|\beta_k| \leq \beta_k^{FR}$, we have that$\cos\theta_k \propto \frac{\|g_k\|}{\|d_k\|}$

Theorem – 2.2:

Suppose that assumption – 2.1 holds and consider any method of the form (2.2), (2.3) where$|\beta_k| \leq \beta_k^{FR}$ and where the step length satisfies the strong Wolf conditions (2.21), (2.22) with $0 < \sigma_1 < \sigma_2 < \frac{1}{2}$, then

$\lim_{k\to\infty} \inf\|g_k\| = 0$

This theorem suggests the following globally convergent modification of the PR method. It differs from that considered by TouatiAhamed and Storey (1990)[34], in that it allows for negative values of$\beta_k, \forall\, k \geq 2$, let$\beta_k = \begin{cases} -\beta_k^{FR}, & if\ \beta_k^{PR} < -\beta_k^{FR} \\ \beta_k^{PR}, & if\ |\beta_k^{PR}| \leq \beta_k^{FR} \\ \beta_k^{FR}, & if\ \beta_k^{PR} > \beta_k^{FR} \end{cases} \longrightarrow$ (2.28)

This strategy avoids one of the main disadvantages of the FR method we will discuss now. By numerical tests that the FR method with in exact line searches sometimes slow down away from the solution: The steps becomes very small and this behavior can continue for a very large number of iterations, unless the method is restarted. This behavior already observed by Powell (1977) [25], who provides an explanation under the assumption of exact line searches. If we extend the argument in the case of inexact line searches, due to (2.26), i.e. suppose that at the iteration k, an unfortunate search direction is generated, such that$\cos\theta_k \approx 0$, and that$x_{k+1} \approx x_k \Longrightarrow \|g_{k+1}\| \approx \|g_k\|\ and\ \beta_{k+1}^{FR} \approx 1 \longrightarrow$ (2.29)

$\Longrightarrow \|g_{k+1}\| \approx \|g_k\| \leq \|d_k\|$, by (2.27) and also by (2.29) and (2.2) we get

$$\|d_{k+1}\| \approx \|d_k\| \geq \|g_{k+1}\| \Longrightarrow \cos\theta_{k+1} \approx 0$$

The argument can therefore start all over again we will give a numerical example later. The PR method would behave quite differently from the FR method in this situation. If$g_{k+1} \approx g_k\ then\ \beta_{k+1}^{PR} \approx 0$, by (2.2) and (2.27) we have$\cos\theta_{k+1} \geq \cos\theta_k \Longrightarrow$The PR method would recover from that situation. Consider the behavior of method (2.28), in this circumstances we have$\beta_{k+1}^{FR} \approx 1\ and\ \beta_{k+1}^{PR} \approx 0$, in this case (3.7) will set$\beta_{k+1} = \beta_{k+1}^{PR}$ as desired and it is reassuring that the modification (2.28) which falls back on the FR method to ensure global convergence avoids the inefficient of this method.

Before we discussed a property of the PR method that is not shared by the FR method, when the step is small,β_k^{PR} it will be small. This property is essential for the analysis of our further research, where a method that possess it, will be said to have Property (*). If the bound$|\beta_k| \leq \beta_k^{PR}$, it can be replaced by$|\beta_k| \leq c\beta_k^{PR}, where\ c > 1$, (it is some suitable constant)$\longrightarrow$ (2.30)

We have not be able to establish global convergence in this case (although, by modifying Lemma 2.1, we can show that the descent property of the search directions can still be obtained provided$\sigma_2 < 1/2c$, we can also see the negative results that is

"Consider the method (2.1) to (2.3) with a line search that always choose the first positive stationary point of$\xi_k(\alpha) = f(x_k + \alpha_k d_k)$, there exists a twice continuously differentiable objective function of three variables, a stationary point and a choice of β_k satisfies$|\beta_k| \leq c\beta_k^{PR}, where\ c > 1$, such that the sequence of gradient$\{\|g_k\|\}$, it is bounded away from zero" Powell (1984), (1985)[26, 27]

Methods related to the PR method with$\boldsymbol{\beta_k \geq 0}$, $\forall$ $\boldsymbol{k}$:

Since PR method cycles without obtaining the solution from Powell examples and also keeping $\beta_k \geq 0$, it is that it allows us to easily enforce the descent property of the algorithm we will discuss here:

Consider the iteration (2.1) to (2.3) with any$\beta_k \geq 0$, we need sufficient condition

$\langle g_k, d_k \rangle \leq -\sigma_3 \|g_k\|^2 for some\ 0 < \sigma_3 \leq 1$ and , $\forall\ k$ ⟶ (2.31)

$\Longrightarrow \langle g_k, d_k \rangle = \|g_k\|^2 + \beta_k \langle g_k, d_{k-1} \rangle$ ⟶ (2.31a)

For the results that follow, we do not specify a particular line search strategy. We assume that the line search satisfies the following three properties

All iterates remain in the level set $\mathcal{L}$ defined in assumption 2.1: $\{x_k\} \subset \mathcal{L}$ ⟶ (2.32)

The Zoutendijk condition holds and the sufficient descent condition (2.31) holds

We already mentioned that the Wolf line search, as well as the ideal line search ensure Zoutendijk condition and reduce f at each step $\Longrightarrow \{x_k\} \subset \mathcal{L}$, and also an exact line search satisfies the sufficient condition (2.31), because$\langle g_k, d_k \rangle \leq -\|g_k\|^2$, i.e. an inexact line search procedure that satisfies the Wolf conditions and (2.31) when$\beta_k \geq 0 \Longrightarrow$ The results apply to both ideal and practical line searches in this situation, for the rest of situation assume that convergence does not occur in a finite number of steps, that is$g_k \neq 0, , \forall\ k$. Further we need some lemmas and theorem for our remaining researches without proof:

Lemma – 2.2:

Suppose that the assumption 2.1 holds. Consider the method (2.1) to (2.3), with$\beta_k \geq 0$, and with any line search satisfying both Zoutendijk condition and sufficient condition (2.31). If $\|g_k\| \geq \gamma, with\ \gamma > 0, \forall\ k$, and then $d_k \neq 0$ $and\ also$ $\sum_{k\geq 2} \|u_k - u_{k-1}\|^2$ $where\ u_k = d_k / \|d_k\|$

This lemma – 2.2 not applicable to the convergence of the sequence$\{u_k\}$ but shows that the search directionu_k, changes slowly and also asymptotically, it applies to any choice of$\beta_k \geq$

0, in addition that β_k it is small when the step is small: That is PR method possess this property and that it prevents the inefficient behavior of the FR method from occurring, that property is essential for our research.

Property (*):

Consider the method of the form (2.1) to (2.3), and suppose that

$$0 < \gamma \leq \|g_k\| \leq \tilde{\gamma} \ \forall\, k \geq 1 \longrightarrow \quad (2.33)$$

Then this method has property (*) if there exist constants $b > 1 \ and \ \lambda > 0 \ such \ that \ \forall\, k$ and then

$$|\beta_k| \leq b \longrightarrow \quad (2.34)$$

$$\text{And} \|s_{k-1}\| \leq \lambda \Longrightarrow |\beta_k| \leq \frac{1}{2b} \longrightarrow \quad (2.35)$$

We can easily verify that under assumption – 2.1, PR and HS methods have property (*)

For the PR method, using constants $\gamma and \tilde{\gamma}$, in (2.33) choose $b = 2\tilde{\gamma}^2/\gamma^2 \, and \lambda = \gamma^2/(2L\tilde{\gamma}b)$, and then we have from (2.5) and (2.33)$|\beta_k^{PR}| \leq \frac{(\|g_k\|+\|g_{k-1}\|)\|g_k\|}{\|g_{k-1}\|^2} \leq 2\tilde{\gamma}^2/\gamma^2 = b$

And when$\|s_{k-1}\| \leq \lambda$, we have from (2.2)$|\beta_k^{PR}| \leq \frac{\|y_{k-1}\|\|g_k\|}{\|g_{k-1}\|^2} \leq L\lambda\tilde{\gamma}/\gamma^2 = 1/2b$

For the HS method, assume that (2.31) and the Wolf second condition are satisfied, and then

$$\langle d_{k-1}, y_{k-1}\rangle = \langle d_{k-1}, g_k\rangle - \langle d_{k-1}, g_{k-1}\rangle \geq -(1-\sigma_2)\langle g_{k-1}, d_{k-1}\rangle \geq (1-\sigma_2)\sigma_3\|g_{k-1}\|^2$$

$$\Longrightarrow \langle d_{k-1}, y_{k-1}\rangle \leq (1-\sigma_2)\sigma_3\gamma^2$$

Again using$\langle d_{k-1}, y_{k-1}\rangle \leq (1-\sigma_2)\sigma_3\gamma^2$, in (2.6), we get$\left|\beta_k^{HS}\right| \leq \frac{2\tilde{\gamma}^2}{(1-\sigma_2)\sigma_3\gamma^2} = b$

Define:$\lambda = (1-\sigma_2)\sigma_3\gamma^2/(2L\tilde{\gamma}b)$ and using$\|g(x) - g(\tilde{x})\| \leq L\|x - \tilde{x}\|$ and if$\|s_{k-1}\| \leq \lambda$, and then

$$\left|\beta_k^{HS}\right| \leq \frac{L\lambda\tilde{\gamma}}{(1-\sigma_2)\sigma_3\gamma^2} = 1/2b$$

We can understand that many other choices ofβ_k give rise to algorithms with Property (*), for example:

If β_k, it has Property (*), so do$|\beta_k| \ and \ \beta_k^+ = \max\{|\beta_k|, 0\}$, again the following lemma – 3 shows that if the gradients are bounded away from zero, and if the method has the Property (*), then a fraction of the steps cannot be too small, first we let $\mathbb{N}^+$(The set of all positive natural numbers) and for $\lambda > 0$, define$\mathcal{K}^\lambda = \{i \in \mathbb{N}^+ | i \geq 2, \|s_{i-1}\| > \lambda\}$, i.e. the set of integers to steps that are larger than λ, and we need to consider groups of Δ consecutive iterates, and for this purpose we define

$\mathcal{K}_{k,\Delta}^\lambda = \{i \in \mathbb{N}^+ | i < k + \Delta - 1, \|s_{i-1}\| > \lambda\}Let\left|\mathcal{K}_{k,\Delta}^\lambda\right|$ denotes the number of elements of$\mathcal{K}_{k,\Delta}^\lambda$

And let $\lfloor . \rfloor$ *and* $\lceil . \rceil$ denote the floor and ceiling respectively

Lemma – 2.3: (without proof)

Suppose that assumption – 2.1 hold and consider the method (2.1) to (2.3) with any line search satisfying $\{x_k\} \subset \mathcal{L}$, and assume that the method has Property (*). Suppose that (2.33) holds, and then there exists $\lambda > 0$ such that for any $\Delta \in \mathbb{N}^+$, and any index k_0, there exists a greater index $k \geq k_0$ such that $\left|\mathcal{K}_{k,\Delta}^{\lambda}\right| > \frac{\Delta}{2}$

Theorem – 2.3: (without proof)

Suppose that assumption – 2.1 hold and consider the method (2.1) to (2.3) with the following properties

$$\beta_k \geq 0, \forall\, k$$

The line search satisfies $\{x_k\} \subset \mathcal{L}$, the Zoutendijk conditions and the sufficient descent conditions with

Property (*) holds

Then $\lim_{k \to \infty} \inf \|g_k\| = 0$

Corollary for theorem 2.3 (without proof):

Suppose that assumption – 2.1 hold and consider the method (2.1) to (2.3) with $\beta_k = \max\{\beta_k^{PR}, 0\}$, and the line search satisfying wolf conditions and the sufficient descent conditions and then $\lim_{k \to \infty} \inf \|g_k\| = 0$

Discussion about our Research: for both Categories:

In Category – 1:

We discussed algorithmic framework that is based on models whose approximation quality is random and sufficiently good with probability $> \frac{1}{2}$, we call such models probabilistically fully linear or fully quadratic depending on the quality of approximation that they provide. Here we discuss how such models can be generated (for some large enough values of the κ constants) and outline future research in this direction.

Fully Linear and Fully Quadratic Polynomial Interpolation Models and Λ– Poised Sample set:

Let $\mathcal{P}_n^d$ denotes the set of polynomials of degree $\leq din \mathbb{R}^n and let q_1 = q + 1$ denotes the dimension of this space. That is the dimension of $\mathcal{P}_n^1 is q_1 = n + 1; for \mathcal{P}_n^2 is q_1 = \frac{1}{2}(n+1)(n+2)$.

A basis $\Phi = \{\phi_0(x), \phi_1(x), \dots, \phi_q(x)\}\, for\, \mathcal{P}_n^d$, it is set of polynomials of degree $\leq d$ that span $\mathcal{P}_n^d$. For any such basis Φ, any polynomial $m(x) \in \mathcal{P}_n^d$, it can be written as

$m(x) = \sum_{j=0}^{q} \alpha_j \phi_j(x)\ where \alpha_j$'s are real coefficients ⟶ (1.4)

Given a set of $p_1 = p + 1\ points\ Y = \{y_0, \dots, y_p\} \subset \mathbb{R}^n, m(x)$, it is said to be the interpolation polynomial of $f(x)\ on\ Y$, if it satisfies

$M(\Phi, Y)\alpha = f(Y)$ ⟶ (1.5)

Where $M(\Phi, Y) = \begin{bmatrix} \phi_0(y^0) & \phi_1(y^0) & \cdots & \phi_q(y^0) \\ \phi_0(y^1) & \phi_1(y^1) & \cdots & \phi_q(y^1) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(y^p) & \phi_1(y^p) & \cdots & \phi_q(y^p) \end{bmatrix}$ ⟶ (1.6)

And $f(Y)$, it is the p_1 dimensional vector whose entries are $f(y_i)\ for\ i = 0, \dots, p$. The interpolation polynomial $m(x)$exists and it is unique if $p = q$, and the set Y, it is poised [30], which essentially means that $M(\Phi, Y)$, it is nonsingular. When the number of pointsp_1, it is smaller than the number of elements in Φ, the matrix $M(\Phi, Y)$, it has more columns than rows and the system (1.5) is underdetermined. In this case, there are several choices of interpolation. If on the other hand, $p > q$ and then the system (1.5), is over determined and we can apply least square regression instead of interpolation. Other polynomial approximations are also possible. If Y, it is such that the condition number of $M(\Phi, Y)$, it is bounded by Λ, where $\hat{Y} = \{(y_0 - x_k)/\Delta, \dots., (y_p - x_k)/\Delta\}$, it is a scalar vector of Y and then Y, it is Λ – poised [8] and also $p_1 > n + 1$, and then we can build a model which is $(\kappa_{ef}, \kappa_{eg})$ – fully linear, with $\kappa_{ef}, \kappa_{eg} = \mathcal{O}(p\Lambda)$ analogously it is shown that if $p_1 > (n + 1)(n + 2)/2$, then we can also build a $(\kappa_{ef}, \kappa_{eg})$ – fully quadratic model, with $\kappa_{ef}, \kappa_{eg} = \mathcal{O}(p\Lambda)$ in n dimensions so we require $(n + 1)(n + 2)/2$ sample points (within reasonable proximity of the current iteratex_k), estimating the condition number of the matrix $M(\Phi, Y)$, it may require $\mathcal{O}(n^6)$ arithmetic operations. This dependency on the dimension limits the use of fully quadratic models to small dimensional problems.

We have two main ways to improve the pre – iteration complexity of DFO algorithms.

One approach is, only change the sample set by one point at a time, has been very successful in practice, as it not only reduce the number of function evaluations, but also the linear algebra involved [35], However in [32] is was shown that such algorithms still require computing a Λ – poised set in the criticality step of the TR framework. Hence the computations of up to n sample points is required if fully linear models are used, while for fully quadratic models up to$(n + 1)(n + 2)/2 - 1$ new sample points have to be evaluated.

The other, complementary, approach is to use quadratic models based on fewer than$(n + 1)(n + 2)/2$ sample points, which also reduces both the cost of the linear algebra

and the number of functions evaluations. In practical DFO applications, incomplete quadratic models have been used very successfully

Let $Y = \{y_0, \dots, y_p\}$, it is a set of sample points with$p < q$, $\Phi = \left\{1, x_1, \dots, x_n, \frac{x_1^2}{2}, x_1x_2, \dots, x_{n-1}x_n \frac{x_n^2}{2}\right\}$

The interpolating polynomial for f on Y is given by (1.4), where α satisfies the undetermined interpolation system (1.6). Since the system admits multiple solutions we have some freedom in selecting α, in [35], the Minimum Frobenius Norm (MFN) models are considered, i.e. the models for which the Frobenius norm of the hessian, or $\|\alpha Q\|_2 = \left\|(\alpha_{n+1}, \alpha_{n+2}, \dots, \alpha_q)\right\|_2$, it is minimized subject to (18). In [30] Powell selects the model based on MFN of the update of the Hessian. Both these methods are successful in practice, and provide useful second order information. However, so far theoretically they are not shown to be superior to simple linear models. Indeed as we will show example below the MFN models may be nearly as bad as simple linear models, but the use of random sample sets can provide a significant practical improvement in this case(see the Table – 1)

Random Samplesets:

In the cases when function evaluations are not very expensive or can be obtained in parallel, there is less incentive to reuse old sample points for the model building, because ensuring the model quality can become the bottleneck of the computations. Instead, we can simply use well – poised deterministic sample set, chosen in advance. However, it is not always the best approach because the pattern is chosen without any consideration for the shape of the function and may be very poor fit. Random sample sets can automatically provide good quality models with high enough probability, yet they do not suffer from the worst case behavior of the deterministic sample sets. (See the Table –1.1)

Table -1.1: The illustration of the oscillating behavior of a TR method based on deterministic underdetermined sample set

Iterations:#	Success	$fvalue$	Δ	ρ
1687	0	$+3.67420711e-02$	$+3.12e-02$	$-1.66e+00$
1688	1	$+3.67418778e-04$	$+6.25e-02$	$+8.13e+02$
1689	0	$+3.67418778e-04$	$+3.12e-02$	$-1.66e+00$
1690	1	$+3.67409693e-04$	$+6.25e-02$	$+3.92e+03$
1691	0	$+3.67409693e-04$	$+3.12e-02$	$-1.66e+00$
1692	1	$+3.67407812e-04$	$+6.25e-02$	$+8.34e+02$
1693	0	$+3.67407812e-04$	$+3.12e-02$	$-1.66e+00$
1694	1	$+3.67398959e-04$	$+6.25e-02$	$+4.13e+03$
1695	0	$+3.67398959e-04$	$+3.12e-02$	$-1.66e+00$

Example of comparison of using undetermined quadratic models based on the random and deterministic sample sets:

Let $f(x) = 10(x_2 - x_1)^2 + (1 - x_2)^2$, it is a version of Rosenbrock function but with smaller curvature and Hessian condition number. Apply the TR method [3] to this function with model based on 5 points at each iteration and construct MFN models based on the sample sets. Note that fully quadratic models require 6 points. Choose deterministic models with the sample set selected as the current iterate plus the coordinate steps of length $\delta, i.e.\ Y = \{y^0, y^0 \pm \delta e_1, y^0 \pm \delta e_2\}$ −a very well poised setin other words the set Yit is generated around the current iteratey^0, by adding coordinate steps of size δ

For the second method we generate the set Y by picking 4 random points in a ball of radius δ around the current iterate. The results are as follows: "The method based on deterministic sample sets achieved the final function value of10^{-4} in 8500 function evaluations, while the method based, while the method based on random sample set achieved (on average over ten runs) the function value of10^{-6} in 2500 function evaluations (with largest deviation of under 200 function evaluations and less that one order of magnitude in accuracy)" Clearly, using random sample sets enhances the performance of the MFN models here. In particular, observe the slow progress of the deterministic method in Table – 2 which represents iteration output. From the Table – 2, the iterations follow a pattern (it starts at around iteration 1000) whereδ_k increased and decreased according to alternating successful and unsuccessful steps, while the progress is slow overall. We repeated experiments 10 times with the fixed pattern which a randomly rotated equivalent of Y defined above. The cyclic behavior and significant slowdown did not occur for every pattern, but for about half of the patterns. The variance in the number of function values was in the order of thousands and the final accuracy varied from 3 to 8 digits, a very non – robust behavior. The purpose of this example is to illustrate the effect of random sample sets. Note that if instead of fixed well – poised deterministic sample sets, use sample sets which included some recent old sample points, then the behavior of MFN method would have matched that of the method based on random sample sets. This is due that fact that the sets of (recent) old sample points have essentially random behavior (Although theoretically this cannot be proved still now)

Analysis of Poisedness of Random Sample Sets:

Let a sample set $Y = \{0, y_0, \dots, y_p\} \subset \mathbb{R}^n$, with a fixed point at the origin and the remaining points being generated randomly from a Standard Gaussian Distribution (With zero mean and covariant matrix equal to the identity matrix, with the case of a scaled identity matrix being a

simple extension of what we present below) and let$\Phi(x) = \{1, x_1, x_2, \dots, x_n\} \Longrightarrow M(\Phi, Y)$, it is a matrix whose first column is all 1's, the first row is zero except the first element and the remaining $p \times n$, it is a Gaussian matrix. Under the simple transformation, the condition number of $M(\Phi, Y) =$the condition number of Random Gaussian $p \times n$ matrix, from the results random matrix theory [11, 12], we got the bound

$P\big(cond(M(\Phi, Y) > \Lambda)\big) \leq C(n,p)\frac{1}{\Lambda^{|n-p|+1}} where C(n,p)$, it is constant dependent on$p and n$

In particular for $p = n$, the result in [11] $\Longrightarrow$ $P\big(cond(M(\Phi, Y) > \Lambda)\big) \leq \frac{1}{\sqrt{2\pi}}\left(\frac{Cn}{\Lambda}\right)$ $where\ C <$ 6.5, it is the universal constant. From this result we get for given $p\ and\ n$ there exists Λ large enough such that

$P\big(cond(M(\Phi, Y) > \Lambda)\big) \geq \frac{1}{2}$. Hence there exists$\kappa_{ef}, \kappa_{eg}$ such that the linear interpolation (or regression) polynomials based on Gaussian sample sets are probabilistically$\big(\kappa_{ef}, \kappa_{eg}\big)$ −fully linear.

Sparse Models based on Random Sample Sets:

For recovery of a sparse fully linear model, if such model exists, in this case the sample set Y, it can be generated by a Gaussian distribution around the current iterate and the random matrix $M(\Phi, Y)$, it can be viewed as a Gaussian matrix just as it described above. Sparse signal recovery can be applied in this well known case to show that if the number of nonzero in the gradient is s, and the number of sample points is $p \geq Cs\log(n/s)$, and then the sparse fully linear model can be recovered with probability greater than

$$1 - c_1 e^{-c_2 p}\ for some universal constants c_1, c_2 and C$$

In fact the constant s also depends on the error between the function values $f(y_i)$, and the sparse values $m(y_i)$, but we omit these details here for simplicity.

Non – uniform Recovery and Martingale Property:

In the example we consider the sample sets are generated to provide high quality of the models independently of the past history of the algorithm. However, our theory allows the probability of a good model to be dependent on the past. In some cases taking this into account may provide a more efficient approach to building models. Here we discuss one possible example. The results of recovery of sparse models which we considered from compressed sensing imply called, uniform recovery, where the matrix $M(\Phi, Y)$, it is designed in such a way that any sparse model can be recovered. However, in our case, it is sufficient to recover the specific model that happens to approximate the objective function f sufficiently well in TR. Thus, the non – uniform recovery results can apply. Some of these results,

including the ones for the Gaussian matrices, can be found in [30a]. The key is that if only one fixed signal needs to be recovered with high probability, and then it is sufficient to generate the random matrix $M(\Phi, Y)$, using fewer samples that what is necessary for the uniform recovery the probability of generating fully linear can be sufficient high, conditioned on the model itself, this fact, in our setting, means that the probability of a "good" model is high, conditioned on the current iterate and TR radius, in other words, on the past behavior of the algorithm. In short, we observe that such a setting will satisfy the sub martingale property, but not complete independence on the past.

Examples of Comparing Performance of Sparse Model recovery vs. Other Underdetermined Second Order Models: Consider the function $f(x) = 10(x_2 - x_1)^2 + (1 - x_2)^2 \; with \; x \in \mathbb{R}^{10}$, it means that we have 10 – dimensional problem, but only the first two dimensions are important. Note that to build a fully linear model without applying sparse recovery we need to sample 11 points, to build a fully quadratic model we need 66 sample points. We apply three variants of the TR algorithm discussed in this paper to this problem which only differ by the choice of the model.

Case – I:

The models are built based on a fixed number$n_Y$of random points that are distributed in a small hypercube around the current iterate, called this method RSTR

Case – II:

We built a sparse models based on "Greedy" sample sets of up to a given number points n_Y, which only use points generated in the course of the TR steps, otherwise, resulting old points, and called it GSTR

Case – III:

This algorithm uses the same greedy sample sets, but construct MFN models, rather than sparse models, Call it MFN

We repeated experiments for each method for$n_Y$ranging from 16 to 40, for RSTR and report results average over 5 runs, since the outcome is random the resulting numbers of function evaluations and iterations are illustrate in Table – 1.2. In the table – 1.2, the number of functions evaluations taken by GSTR and MFN are roughly same as the number of iterations, because only one function value is computed per iteration, except for the first iteration. From the number of iterations required by each algorithm it is evident that RSTR clearly recovers the fully quadratic models of f , with as few as 20 sample points, while the other two methods do not. While the first algorithm performs more function evaluations, they can be

obtained in parallel, and the achieved accuracy is by far better than that of the other methods. Note that while the outcome of RSTR is random, we observed very little variance in the number of iterations, and hence, function evaluations, in our experiments. It is also iterating to observe that the number of iterations of RSTR does not get smaller after the size of the $n_Y$exceeds 20, hence, it appears that while larger number of sample points per iteration may allow for more reliable recovery, it is not necessary for fast convergence. Computational traded – offs in optimization using sparse models is a subject of a separate study.

Table – 1.2: The comparison of the number of iterations and function evaluation for RSTR, averaged over 5 runs,

(Accuracy for$n_Y \geq 20$ is $1.0e-11$), GSTR (accuracy$1.0e-11$) and MFN (accuracy$1.0e-5$) methods

n_Y	# Of iter. RSTR	# Of iter. GSTR	# Of iter. MFN	# Of iter. Eval. RSTR
16	189	318	767	3027
17	99	247	1048	1693
18	38	158	925	680
19	30	157	646	585
20	27	137	1086	536
21	20	144	919	424
22	21	145	778	462
23	24	149	848	547
24	19	150	962	515
25	20	152	636	504
26	19	144	414	513
27	19	125	456	509
28	18	166	401	568
29	20	137	659	570
30	19	122	503	607
31	20	154	349	627
32	20	157	390	646
33	20	164	484	528
34	18	175	442	605
35	18	164	362	637
36	20	139	301	727
37	20	162	328	754
38	20	136	439	652
39	18	183	316	702
40	19	152	536	760

Discussion about our Research:In Category – 2:

We saw that global convergence is obtained for any β_k, in the interval$I_1 = [-\beta_k^{PR}, \beta_k^{FR}]$, and we proved global convergence for any β_k, with Property (*) contained in the interval$I_2 = [0, \infty]$. Whether these results can be combined to obtain larger intervals of admissible β_k: In

particular, since PR method has property (*), whether global convergence is obtained by restrictingβ_k^{PR}, to the larger interval$I_1 \cup I_2$, i.e. by letting $\beta_k = \begin{cases} \beta_k^{PR}, \ if\ \beta_k^{PR} \geq \beta_k^{FR} \\ -\beta_k^{PR}, \ otherwise \end{cases}$

It is enough, global convergence cannot be guaranteed, and it was shown by Powell (1984) and the sequence$\{\beta_k^{PR}, \beta_k^{FR}\}$, it has exactly three accumulation points $-\frac{1}{3}, 1\ and\ 10$

$\Longrightarrow \exists$, an integerk_0 such that $\beta_k = \beta_k^{PR} \geq \beta_k^{FR}, \forall\ k \geq k_0$, and now, the function can be modified and the starting point can be changed so that PR method generates from the new initial point$\tilde{x}_1$, a sequence$\{\tilde{x}_k\}\ with\ \tilde{x}_k = x_{k+k_{0-2}}\ for\ k \geq 2$. In this modified case, we have$\tilde{\beta}_k^{PR} \geq -\tilde{\beta}_k^{FR}, \forall\ k \geq 2$, but the sequence of gradient is bounded away from zero.We check another case in which interval of admissible β_k, it cannot be combined. Any method of the form (2.2) to (2.3) with a line search giving$\langle g_k, d_{k-1} \rangle = 0, \forall\ k, and\ with\ \beta_k \in I_3 = [-1, 1]$, it is globally convergent. This is easy to see, since in this case$\|d_k\|^2 \leq \|g_k\|^2 + \|d_{k-1}\|^2 \leq \cdots \leq \tilde{\gamma}^2 k, where\ \tilde{\gamma}$, it is an upper bound on

$\|g(x)\| \Longrightarrow \|d_k\|^2$, grows at most linearly, and globally convergence (explained in preliminaries). Otherwise, by corollary of $(theo.2.3)$ shows that the PR method is convergent if restricted to$I_1 = [0, \infty)$.

However the PR method may not converge ifβ_k^{PR}, it is restricted to$I_3 \cup I_2 = [-1, \infty)$. This argument again based on the fact that$\beta_k^{PR} \geq -\frac{1}{4}, \forall\ k$ (It is proved by means of Cauchy – Schwartz inequality).

$\Longrightarrow$, in the case of$\beta_k^{PR} \in [-1, \infty)$, but convergence is not obtained. Therefore we are not able to generalize in FR method, PR method with $\beta_k \geq 0$, and we look more closed at the conditions used in these cases. First we check under what conditions is$\beta_k^{PR} \geq 0\ \ or\ \beta_k^{PR} \geq -\beta_k^{FR}$. For strictly convex quadratic functions and exact linear searches, the PR method coincides with the FR method. Since always$\beta_k^{FR} > 0\ so\ is\ \beta_k^{PR}$, let us consider strongly convex functions with $(\beta_k^{PR} < 0, \text{and} \beta_k^{PR} \leq -\beta_k^{FR})$

We will prepare one important Proposition – 1:

Proposition – 1:

There exists a$\mathbb{C}^\infty$ strongly convex function of two variables and a stationary pointx_1, for which PR method with exact line searching gives$\beta_3^{PR} < -\beta_3^{FR} < 0$

Proof:

Introduce strictly convex quadratic function $\tilde{f}$ of two variables $x = (x_{(1)}, x_{(2)})$ as $\tilde{f}(x) = x_{(1)}^2 + \frac{1}{2}x_{(2)}^2$ with gradient and Hessian (the Euclidean Scalar Product is assumed), i.e. $\nabla\tilde{f}(x) = \begin{bmatrix} 2x_{(1)} \\ x_{(2)} \end{bmatrix}$; $\nabla^2\tilde{f}(x) = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$

Starting from the point $x_1 = (-3, 3)$, the PR method with exact line searches gives

$\nabla\tilde{f}(x_1) = \begin{bmatrix} -6 \\ 3 \end{bmatrix}, \tilde{\alpha}_1 = \frac{5}{9}$ and $\tilde{x}_2 = \frac{1}{3}\begin{bmatrix} 1 \\ 4 \end{bmatrix}$, next it finds

$$\nabla\tilde{f}(\tilde{x}_2) = \frac{2}{3}\begin{bmatrix} 1 \\ 2 \end{bmatrix}. \beta_2^{PR} = \frac{4}{81}, \tilde{d}_2 = -\frac{10}{27}\begin{bmatrix} 1 \\ 4 \end{bmatrix} \; and \; \tilde{\alpha}_2 = \frac{9}{10}$$

The third point is the solution point $x_* = (0, 0)$ now perturb the function $\tilde{f}$ inside the ball

$B(0, 1) = \left\{x | x_{(1)}^2 + \frac{1}{2}x_{(2)}^2 < 1\right\}$

Defining $f(x) = \tilde{f}(x) + \varepsilon\psi(x)$ $where$ $\varepsilon > 0$ $is\ small, \psi\ such\ that$

$\psi(x) = 0, \forall\ x \notin B(0, 1)$ ⟶ (1.7)

As the line joining x_1 and $\tilde{x}_2$, it does not intersect the closure of $B(0, 1)$, i.e. the PR method on the new function, starting from the same point x_1, gives $x_2 = \tilde{x}_2$ and $d_2 = \tilde{d}_2$

To prove how to choose the function ψ and $\varepsilon > 0$, so that f, it is strongly convex and $\beta_3^{PR} < 0$

Take $\psi(x) = \eta(x)\ell(x)$ $where$ $\ell = 4x_{(1)} - x_{(2)}$, it is a linear function and $\eta(x)$ in $\mathbb{C}^\infty$ satisfying

$$\eta(x) = \begin{cases} 1, & if\ x \in B\left(0, \frac{1}{2}\right) \\ 0, & if\ x \notin B(0, 1) \end{cases}$$

Clearly ψ satisfies (1.7) and it has bounded second order derivatives $\Longrightarrow$ Choosing ε sufficiently small, say, $0 < \varepsilon < \varepsilon_1$, and then Hessian of f, it will be uniformly positive definite and f in $\mathbb{C}^\infty$, strongly convex function. Now f, it is determined in this manner, there is a unique minimum of f, from x_2, in the direction d_2, as $\nabla f(0) = \nabla\tilde{f}(0) + \varepsilon\nabla\psi(0) = \varepsilon\begin{bmatrix} 4 \\ -1 \end{bmatrix}$, it is orthogonal to $d_2 = \tilde{d}_2$, the one – dimensional minimum is still is obtained at $x_3 = (0, 0)$ (But this is no longer solution point)

$\Longrightarrow \beta_3^{PR} + \beta_3^{FR} = \frac{2|\nabla f(0)|^2 - \langle\nabla f(0), \nabla f(x_2)\rangle}{|\nabla f(x_2)|^2} = \frac{34\varepsilon^2 - (4\varepsilon/3)}{20/9}$. That is $\beta_3^{PR} < -\beta_3^{FR} < 0, if\ 0 < \varepsilon < \varepsilon_2 = \frac{2}{51}$

By taking $\varepsilon \in (0, \min\{\varepsilon_1, \varepsilon_2\})$, we obtain the desired result. This proposition – 1 show that the convergence result, which was obtained for strongly convex function and exact line –

searches, is not a consequence of theorem – 2.3, since the latter requires $\beta_k \geq 0$. Nor is it a consequence of theorem – 2.2, because the proposition – 1 shows thatβ_k^{PR}, it can lie outside the interval$[-\beta_k^{FR}, \beta_k^{PR}]$

Numerical Experiments:

We have tested several of the algorithms suggested by the convergence analysis of this paper, on the collection largest test problems given in Table –2.3: The starting points used are those given in the references. For example: The problems of $Moor\ et.al$[22] we set the parameter factor $= 1$, for test problems 8, 9 and 10 starting point 3 from the reference was used. We verify that each run, all the methods converge to the same solution point; otherwise the problem was not included in test set. The problems are not numbered consecutively because they belong to a larger test set. Since the Conjugate Gradient Methods are mainly used for large problems, our test problems have at least 100 variables.

The following are the methods tested; they differ, only the choice of β_k and possibly, in the line search.

1. FR⟶The Fletcher – Reeves Method
2. PR – FR⟶The Polack – Ribiere Method construct by the PR method (shown before)
3. PR⟶The Polack – Ribiere Method
4. PR^+ ⟶The Polack – Ribiere Method allowing only $\beta_k > 0$

For the line – search we used the algorithm of $Moor\ \ and\ Thuente\ et.al$ (1990)[22]. This algorithm finds a point satisfying the strong Wolf Conditions. We used the values$\sigma_1 = 10^{-4}\ \ and\ \sigma_2 = 0.1$, which by theorem – 2.2, ensure that methods FR and PR – FR are globally convergent. The line – search for PR andPR^+ methods was performed as follows: First find a point satisfying the Strong Wolf Conditions, using the values of$\sigma_1\ and\ \sigma_2$ mentioned as above. If at this point the directional derivative of $f < 0$, we know that the sufficient descent condition holds, and terminate the line search (already discussed before). On the other hand, if the directional derivative of $f > 0$, the algorithms of $Moor\ \ and\ Thuente$, it has bracketed a one – dimensional minimizer, and if the line search iteration is continued it will give, in the limit, a point $x_k\ \ with\ \langle g_k, d_{k-1}\rangle = 0$. By continuity and (2.31a), it is clear that the line search will find a point satisfying the sufficient descent condition in a finite number of iterations. In the numerical we set$\sigma_3 = 10^{-2}$, in the sufficient descent condition.Our numerical experience with conjugate gradient methods indicates that it is advantageous to perform a reasonably accurate line search. Therefore in addition to setting$\sigma_2 = 0.1$, we ensured that the line search evaluated the function at least

twice. The choice of the initial trial line search is also important. For the first iteration set it to$1/\|g_1\|$, and for subsequent iterations we used the formula recommended byShanno and Phua (1980) [31], which is based on quadratic interpolation.The tests were performed on SPRC station – 1, using OCTAVE in double precision. All the runs were stopped when$\|g_k\|_\infty < 10^{-5}(1 + |f(x_k)|)$. The results in Table – 2.2 and Table – 2.3 are given in the form: (Number of iterations)/ (Number of function evaluation). The number under the column "mod" for method PR – FR denotes the number of iterations for which$|\beta_k^{PR}| > \beta_k^{FR}$: For$PR^+$, "mod" denotes the number of iterations for which$\beta_k^{PR} < 0$. If the limit of 9999 function evaluations was exceeded the run was stopped; this is indicated by "*".This generally occurs when the stopping criterion is very demanding.

Table –2.1: List of Test Functions for Numerical Experiments

Problems	Name	References	n
2	Calculus of Variations 2	Gill and Murray (1973)[15]	100, 200
3	Calculus of Variations 3	Gill and Murray (1973)[15]	100, 200
6	Generalized Rosenbrock	Moor et al. (1981)[21]	100, 500
8	Penalty 1	Gill and Murray (1979)[4]	100, 1000
9	Penalty 2	Gill and Murray (1979)[16]	100
10	Penalty 3	Gill and Murray (1979)[16]	100, 1000
28	Extended Powell Singular	Moor et al. (1981) [21]	100, 1000
31	Brown almost linear	Moor et al. (1981)[21]	100, 200
38	Tri diagonal 1	Buckley and LeNir (1983) [3]	100, 1000
39	Linear minimal surface	Toint (1983)[33]	121, 961
40	Boundary – value problem	Toint (1983)[33]	100
41	Broyden tri diagonal nonlinear	Toint (1983)[33]	100
42	Extended ENGVL1	Toint (1983)[33]	100, 10000
43	Extended Freudenstein and Roth	Toint (1983)[33]	100, 1000
45	Wrong extended Wood	Toint (1983)[33]	100
46 (1)	Matrix square root$(ns = 1)$	Liu and Nocedal (1988)[19]	100
46 (2)	Matrix square root$(ns = 2)$	Liu and Nocedal (1988)[19]	100
47	Sparse matrix square root	Liu and Nocedal (1988)[19]	100, 1000
48	Extended Rosenbrock	Moor et al. (1981)[21]	1000, 10000
49	Extended Powell	Moor et al. (1981)[21]	100, 1000
50	Tri diagonal 2	Toint (1983)[33]	100, 1000
51	Trigonometric	Moor et al. (1981)[21]	100, 1000
52	Penalty 1 (2nd version)	Moor et al. (1981)[21]	1000, 10000

Table – 2.2: Smaller Problems

PN		FR	PR - FR		PR	PR^+	
		$it/f-g$	$it/f-g$	mod	$it/f-g$	$it/f-g$	Mod
2100		405/827	405/820	351	400/812	405/812	0
3	100	1313/2627	1313/2627	1313	1299/2599	1299/2599	0
6	100	*	261/547	95	256/529	254/525	1
8	100	10/36	15/49	12	9/39	12/47	2
9	100	7/20	8/22	6	8/25	7/20	2
10	100	116/236	93/191	91	118/244	119/244	1
28	100	1426/2855	1291/2584	1289	120/280	168/382	3
31	100	2/3	2/3	1	1/4	1/4	0

38	100	70/142	70/142	47	71/144	71/144	0
39	121	*	59/122	4	59/122	59/122	0
40	100	175/351	175/351	175	132/266	132/266	0
41	100	29/60	24/50	1	24/50	24/50	0
42	1000	10/27	9/25	8	10/34	9/30	2
43	100	16/41	14/39	13	16/44	13/37	1
45	100	*	74/166	66	37/90	45/109	3
46(1)	100	617/1238	253/510	248	257/518	257/518	0
46(2)	100	886/1776	251/506	243	251/506	251/506	0
47	100	151/306	59/122	50	60/124	60/124	0
48	1000	79/185	71/172	66	26/73	23/70	3
49	100	1426/2855	1291/2584	1289	117/281	168/382	3
50	100	72/146	72/146	52	72/146	72/146	0
51	100	202/409	42/94	12	45/103	45/103	0
52	1000	3/10	3/10	2	4/12	4/12	2

Table – 2.3: Larger Problems

PN		FR	PR – FR		PR	PR^{+}	
		$it/f-g$	$it/f-g$	Mod	$it/f-g$	$it/f-g$	mod
2200		703/1424	701/1420	591	701/1420	701/1420	0
3	200	2808/5617	2808/5617	2808	2631/5263	2631/5263	0
6	500	*	1107/2231	433	1068/2151	1067/2149	1
8	1000	12/39	9/34	7	6/28	10/42	2
10	1000	138/281	145/299	142	165/338	165/338	0
28	1000	533/1102	1369/2741	1366	212/473	97/229	3
31	200	2/4	2/4	1	1/5	1/5	0
38	1000	264/531	263/529	217	262/527	262/527	0
39	961	*	143/220	5	142/287	142/287	0
42	10000	6/26	6/26	5	7/28	6/26	1
43	1000	10/27	15/38	15	10/33	9/29	2
47	1000	422/849	114/233	92	113/231	113/231	0
48	10000	61/143	130/283	123	24/73	19/62	4
49	1000	568/1175	1369/2741	1366	212/473	97/229	3
50	1000	274/551	273/549	245	274/551	274/551	0
51	1000	231/467	40/91	5	40/92	40/92	0
52	10000	4/15	4/15	4	3/13	3/13	1

Conclusion and Future Study for Global Convergence:

For Category – 1:Other Random Models:

Additional settings where relying on random models may given an advantage for an optimization scheme occur in a parallel environment when full synchronization is not needed: We may wish to consider the following asynchronous setting: Each processor takes, differ random amount of time to compute a function value. Let us consider a time budget τ, and assume that a sufficiently large number of function evaluations are computed in less than τ time, with some sufficiently high probability. If we assign function evaluations to processors randomly, then the resulting sample set is random and the resulting model is well

– poised with high probability. Alternatively we may consider a setting the objective function is evaluated approximately for each sample point, with some high probability of this approximation being accurate, but yet some small probability of a bad approximation. In this case the resulting interpolation/Regression model will provide a good approximation with high probability. Note that when computing the function value at the potential new iterate (rather than a sample point) and assuming that an accurate value is computed. Relaxing this condition is also a subject for future study.

Reusing Sample Points:

In sequential computational setting with expensive function evaluations it is efficient to reuse existing sample points in the vicinity of the current iterations. The success of the second method in the example above indicates that sparse models based on greedy sample sets are useful, even though the sparse recovery properties are unlikely to hold for such sets. Hence the random sample models may be dependent in some practical approaches. An obvious example o a method which reuses sample points and relies on sample models would be as follows. For each of iterations the interpolation model is build based on up to $n+1$ existing sample points and an additional number of random sample points. The past points picked so that they are in a reasonable vicinity of the current iterate and so that they form a well – poised set. For various technique of selecting such a set [8], the random points are selected to enrich the current set of sample points. The resulting sample set clearly depends on the history of the algorithm, on the other hand, if the set of re – used sample points is well poised, then the whole set is well – poised with sufficiently high probability. Investigating general cases of models when the sub martingale property holds, or relaxing the sub martingale property in a controlled way, and deriving new convergence results is a subject of our future research. Before more detailed analysis is derived it is essential to identify classes of random models that best perform in practice.

For Category – 2:

From the table 2.1 to 2.3, we conclude thatβ_k^{PR}, it was constrained in most of the iterations of the method PR – FR, but was quite rarely modified in thePR^+ method. Many of the problems were run again for a larger number of variables. The results are given in Table – 3: The performance of methods Pr – FR, PR andPR^+, it is comparable. Over all, PR^+ appears to be better that PR. The FR method is clearly the last efficient, requiring an exceedingly large number of function evaluations in some problems.

In these runs the methods were implemented without restarting. We also performed test in which the methods were restarted along the steepest descent direction every n iterations. (Since n it is large, very few restarts were performed). The FR method improved substantially, but this method was still the least efficient of the four. The other three methods performed similarly with and without restarts, and we will not present the results here: We can give an example that illustrates the inefficient behavior of the FR method, as mentioned before. In the Table Problem – 45 with $n = 100$, observed that for hundreds of iterations$\cos\theta_k$ stays fairly constant, and is of order10^{-2}, while the steps$\|x_k - x_{k-1}\|$, they are of order10^{-2} to 10^{-3}. This causes the algorithm to require a very large number of iterations to approach the solution. A restart along the steepest descent direction terminates this cycle of bad search directions and tiny steps. A similar behavior was observed in several other given problems in Table – 2.1.

REFERENCES

1. Al – Baali, M. "Descent property and global convergence of the Fletcher – Reeves method with inexact line search" *IMA Journal of Numerical Analysis*, Vol. 5, pp. 121–124, (1985).
2. Billups, S.C. Larson, J. and Graf, P. "Derivative – Free – Optimization of expensive functions with computational error using weighted regression" *SIAM J. Optim*, Vol. 23. Pp. 27 – 53, (2013).
3. Buckley, A. and LeNir, A. "QN – like variable storage conjugate gradients", *Mathematical Programming*, Vol. 27, pp. 367 – 175, (1983).
4. Conn, A.R. Scheinberg, K and Tiont, Ph. L. "Trust Region Methods" *MPS – SIAM series on Optimization*. *SIAM, Philadelphia*, (2000).
5. Conn, A.R. Scheinberg, K and Tiont, Ph. L. "On the convergence of Derivative – Free Methods for unconstrained optimization" *In M.D. Buhuman and A Iserels, editors, Approximation Theory and Optimization, Tributes to M.J.D. Powell* pages 83 – 108, Cambridge University Press, Cambridge, (1997).
6. Conn, A.R. Scheinberg, K and Tiont, Ph. L. "Recent progress in unconstrained nonlinear optimization without derivatives", *Math. Program*, 79:397 – 414, (1997).
7. Conn, A.R. Scheinberg, K and Vincent L.N."Geometry of sample sets in derivative free optimization polynomial regression and underdetermined interpolation". *Math. Program*. 111: 141 – 172, (2008).

8. Conn, A.R. Scheinberg, K and Vincent L.N. "Global convergence of general derivative – free trust – region algorithms to first and second order critical points" *SIAM J. Optim.* 20: 387 – 415, (2009).
9. Conn, A.R. Scheinberg, K and Vincent L.N. "Introduction to Derivative – free Optimization" *SIAM, Philadelphia,* (2010).
10. Durrett, R. Probability: "*Theory and Examples. Cambridge Series in Statistical and Probabilistic Mathematics*" Cambridge, fourth edition, (2010).
11. Edelman, A. "Eigen Values and condition numbers of random matrices" *SIAM J. Matrix Anal. Appl.* 9: 543 – 560, (1998).
12. Fasano, G. Morales, J.L. and Nocedal, J. "On the geometry phase in model – based algorithms for derivative – free optimization" *Optim. Methods Soft.* 24: 145 -154, (2009).
13. Fletcher, R. and Reeves, C. M "Function minimization by conjugate gradients" *Computer Journal* Vol. 7, pp. 149 – 154. (1964).
14. Fletcher, R. "*Practical Methods of Optimization*" Wiley, (1987).
15. Gill, P.E. and Murray, W. "The Numerical solution of a problem in the Calculus of Variations, in D.J. Bell, ed., *Recent Mathematical Development on Control,* Academic Press, New York, pp: 97 – 122 (1973).
16. Gill, P.E. and Murray, W. "Conjugate – gradient methods for large – scale nonlinear optimization, Technical report SOL 79 – 15", *Department of Operation Research,* Stanford University, Stanford, CA (1979).
17. Gill, P.E. and Murray, W. and Wright, M. H. "Practical Academic Press. (1981)
18. Kolda, T.G. Lewis, R.M. and Torczon, V. "Optimization by direct search: New perspectives on some classical and modern methods" SIAM Rev. Vol. 45, pp. 385 – 482, (2003)
19. Liu, D.C. and Nocedal, J. "Test results of two limited memory methods for large sacle optimization, Report NAM 04" *Department of Electrical Engineering and Computer Science,* Northwestern University. (1988).
20. Matyas, J. "Random optimization" *Automation and Remote Control,* 26: 240 – 253, (1965).
21. Moor, J. J. Garbow, B.S. and Hillstrom, K. E. "Testing unconstrained optimization software ACM" *Transactions on Mathematical Software,* Vol. 7, pp: 17 – 41, (1981).

22. Moor, J.J. and Thuente, D. J. "Online search algorithms with guaranteed sufficient decrease" Mathematics and Computer Science Division Preprint MCS – P153 – 0590", Argonne *National Laboratory*, Argonne II 60439. (1990)
23. Moor, J.J. and Wild, S. M. "Benchmarking derivative – free optimization algorithms" SIAM J. Optim. 26: 172 – 191, (2009).
24. Polak, E. and Ribiere, G. "Note sur la convergence de methods de directions conjugates, Revue Francaised'Informatique et.edRechercheOpterationnelle" Vol. 16. Pp. 35 – 43, (1969).
25. Powell, M.J.D. "Restart proceduce of the conjugate gradient methods" *Mathematical Programming*, Vol. 12. pp. 241 – 254, (1977).
26. Powell, M.J.D. "Non – convex minimization calculations and the conjugate gradient methods, in *Lecture Notes in Mathematics*, Vol. 1066, pp. 122 – 141, Springer – Verlag, Berlin. (1984).
27. Powell, M.J.D. "Convergence properties of algorithms for nonlinear optimization, Report DAMTP 1985/NA1" Department of Applied Mathematics and Theoretical Physics, University of Cambridge, England, *Presented at the 1984 SIAM Summer Meeting in Seattle*, (1985)
28. Powell, M.J.D. "A direct search optimization method that models the objective and constraint functions by linear interpolation. In S. Gomez and J.P. Hennart, editors" *Advances in Optimization and Numerical Analysis, Proceedings of Sixth, Workshop on Optimization and Numerical Analysis, Oaxaca, Mexico.* Vol. 275 of Mat. Appl. Pp: 51 – 67. Kluwer Academic Publishers, Dordrecht, (1994).
29. Powell, M.J.D. "On trust region methods for unconstrained minimization without derivatives" *Mathematical Programming*, Vol. 97, pp: 605 – 623, (2003)
30. Powell, M.J.D. "Least Frobenius norm updating of quadratic models that satisfy interpolation conditions" *Mathematical Programming*, Vol. 100, pp. 183 – 215, (2004)

30 Rauhut, H. "Compressive sensing and structured random matrices. In M. Fornasier, editor" Theoretical *Foundation and Numerical Methods for Sparse Recovery, Radon Series Comp*. Appl. Maths, pages 1 – 92 (2010)

31. Shanno, D.F. and Phua, K. H. "Remark on algorithm 500: minimization of unconstrained multivariate functions" ACM Transactions on Mathematical Software, Vol. 6, pp. 618 – 622, (1980)

32. Scheinberig, K. and Toint, Ph. L. "Self – correction geometry in model – based algorithms for derivative – free unconstrained optimization" SIAM J. Optim. Vol. 20, pp. 3512 – 3532, (2010)
33. Toint, Ph. L. "Test problems for partially separable optimization and results for the routine PSPMIN, Report 83/4" *Department of mathematics, Faculties Universitaire* de Namur, Namur, Belgium, (1983).
34. Touati – Ahmed, D. and Storey, C. " Efficient hybrid conjugate gradient techniques", Journal of Optimization Theory and Applications, Vol. 64, pp. 379 – 397, (1990)
35. Wild, S. M. MNH: "A derivative – free optimization algorithm using minimal norm Hessians" In Tenth Copper Mountain Conference on Iterative Methods, April, (2008)
36. Wolf, P. "Convergence conditions for ascent methods", *SIAM Review*, Vol. 11, pp, 226 – 235, (1969).
37. Wolf, P. "Convergence conditions for ascent methods – II some corrections", *SIAM Review*, Vol. 13, pp, 185 – 188, (1971).
38. Zoutendijk, G. "Nonlinear Programming, Computational methods" *in Integer and Nonlinear Programming* pp. 37 – 86, ed. J Abadie, North – Holland, Amsterdam.