

Q-fully quadratic modeling and its application in a random subspace derivative-free method

Yiwen Chen

Department of Mathematics
University of British Columbia

July 25, 2024

Joint work with Dr. Warren Hare and Dr. Amy Wiebe

- 1 Introduction
- 2 Quadratic approximation random subspace trust-region algorithm
- 3 Convergence analysis
- 4 Numerical experiments
- 5 Summary

Derivative-free optimization (DFO)

Consider the optimization problem

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

where f is given by a blackbox:



Derivative-free optimization is the mathematical study of optimization algorithms that do not use derivatives

Note: It does not mean that the derivatives do not exist

Model-based DFO methods:

- Use function values to build an approximation model of the objective
- Use the model to guide future iterations

Limitations:

- Function evaluations are too expensive for large problems
- They are primarily designed for small- to medium-scale problems

Improving scalability by subspace decomposition

Idea:

1. Select a low-dimensional affine subspace
2. Build and minimize a model to compute a step in this subspace
3. Change the affine subspace at the next iteration

Some existing papers:

[Zhang, 2012]; [Dzahini, Wild, 2022]; [Menickelly, 2023];
[Cartis, Roberts, 2023]

Outline

- 1 Introduction
- 2 Quadratic approximation random subspace trust-region algorithm
- 3 Convergence analysis
- 4 Numerical experiments
- 5 Summary

Model-based trust-region (MBTR) algorithm

for $k = 0, 1, \dots$ **do**

Construct a model m_k in \mathbb{R}^n :

$$m_k(s) = f(x_k) + g_k^\top s + \frac{1}{2} s^\top H_k s$$

Approximately solve the trust-region subproblem in \mathbb{R}^n :

$$s_k \approx \underset{s \in \mathbb{R}^n}{\operatorname{argmin}} m_k(s), \quad \text{s.t. } \|s\| \leq \Delta_k$$

Evaluate $f(x_k + s_k)$ and calculate ratio

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(\mathbf{0}) - m_k(s_k)} = \frac{\text{true decrease}}{\text{predicted decrease}}$$

Accept/reject step based on ρ_k and update trust region radius

Random subspace MBTR algorithm

for $k = 0, 1, \dots$ **do**

Define an affine subspace $x_k + D_k \mathbb{R}^p$ by selecting $D_k \in \mathbb{R}^{n \times p}$

Construct a model \widehat{m}_k in \mathbb{R}^p

Approximately solve the trust-region subproblem in \mathbb{R}^p :

$$\widehat{s}_k \approx \underset{\widehat{s} \in \mathbb{R}^p}{\operatorname{argmin}} \widehat{m}_k(\widehat{s}), \quad \text{s.t. } \|\widehat{s}\| \leq \Delta_k$$

and calculate the corresponding step $s_k \in \mathbb{R}^n$

Evaluate $f(x_k + s_k)$ and calculate ratio

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{\widehat{m}_k(\mathbf{0}) - \widehat{m}_k(\widehat{s}_k)} = \frac{\text{true decrease}}{\text{predicted decrease}}$$

Accept/reject step based on ρ_k and update trust region radius

Model construction

Q-fully quadratic models

Definition. Let $f \in \mathcal{C}^2$, $x \in \mathbb{R}^n$, $\bar{\Delta} > 0$, and $Q \in \mathbb{R}^{n \times p}$

We say that $\{\widehat{m}_\Delta : \mathbb{R}^p \rightarrow \mathbb{R}\}_{\Delta \in (0, \bar{\Delta}]}$ is a class of Q-fully quadratic models of f at x if there exist $\kappa_f(x), \kappa_g(x), \kappa_h(x) > 0$ such that for all $\Delta \in (0, \bar{\Delta}]$ and $\|\widehat{s}\| \leq \Delta$,

$$\begin{aligned} |f(x + Q\widehat{s}) - \widehat{m}_\Delta(\widehat{s})| &\leq \kappa_f(x)\Delta^3, \\ \left\| Q^\top \nabla f(x + Q\widehat{s}) - \nabla \widehat{m}_\Delta(\widehat{s}) \right\| &\leq \kappa_g(x)\Delta^2, \\ \left\| Q^\top \nabla^2 f(x + Q\widehat{s}) Q - \nabla^2 \widehat{m}_\Delta(\widehat{s}) \right\| &\leq \kappa_h(x)\Delta \end{aligned}$$

Generalized simplex gradient and Hessian

Definition. [Hare, Jarry-Bolduc, 2020][Hare, Jarry-Bolduc, Planiden, 2023]

Let $x^0 \in \mathbb{R}^n$ and $D = [d_1 \cdots d_p] \in \mathbb{R}^{n \times p}$

The **generalized simplex gradient** of f at x^0 over D is defined by

$$\nabla_S f(x^0; D) = (D^\top)^\dagger \delta_f(x^0; D)$$

The **generalized simplex Hessian** of f at x^0 over D is defined by

$$\nabla_S^2 f(x^0; D) = (D^\top)^\dagger \delta_{\nabla_S f}(x^0; D)$$

Constructing Q -fully quadratic models

Theorem. Suppose $D = QR$ has full column rank, where $Q \in \mathbb{R}^{n \times p}$
Let

$$m(x) = f(x^0) + (2\nabla_S f(x^0; D) - \nabla_S f(x^0; 2D))^T (x - x^0) \\ + \frac{1}{2} (x - x^0)^T \nabla_S^2 f(x^0; D) (x - x^0)$$

Define model $\widehat{m} : \mathbb{R}^p \rightarrow \mathbb{R}$ by

$$\widehat{m}(\widehat{s}) = m(x^0 + Q\widehat{s})$$

If $f \in \mathcal{C}^{2+}$ and D has full-column rank, then \widehat{m} belongs to a class of Q -fully quadratic models of f at x^0 with constants $\kappa_f, \kappa_g, \kappa_h$ monotonically increasing w.r.t. $\|D^\dagger\|$

Subspace selection

Definition. [Cartis, Roberts, 2023]

Let $f \in \mathcal{C}^1$, $x \in \mathbb{R}^n$, and $\alpha > 0$

We say that $A \in \mathbb{R}^{n \times z}$ is α -well aligned for f at x if

$$\left\| A^T \nabla f(x) \right\| \geq \alpha \|\nabla f(x)\|$$

Constructing α -well aligned matrices

Theorem. [Dzahini, Wild, 2022]

Suppose $\alpha, \delta_S \in (0, 1)$ and $z \geq 4(1 - \alpha)^{-2} \ln(1/\delta_S)$

Let $A \in \mathbb{R}^{n \times z}$ such that $A_{ij} \sim \mathcal{N}(0, 1/z)$

Then for any $v \in \mathbb{R}^n$,

$$\mathbb{P} \left[\left\| A^\top v \right\| \geq \alpha \|v\| \right] \geq 1 - \delta_S$$

The $\mathbb{P}[\cdot]$ gives the probability of a random variable

Constructing α -well aligned matrices

Theorem. [Dzahini, Wild, 2022]

Suppose $\alpha, \delta_S \in (0, 1)$ and $z \geq 4(1 - \alpha)^{-2} \ln(1/\delta_S)$

Let $A \in \mathbb{R}^{n \times z}$ such that $A_{ij} \sim \mathcal{N}(0, 1/z)$

Then for any $v \in \mathbb{R}^n$,

$$\mathbb{P} \left[\left\| A^\top v \right\| \geq \alpha \|v\| \right] \geq 1 - \delta_S$$

In particular, given $f \in \mathcal{C}^1$ and $x \in \mathbb{R}^n$, A is α -well aligned for f at x with probability at least $1 - \delta_S$, i.e.,

$$\mathbb{P} \left[\left\| A^\top \nabla f(x) \right\| \geq \alpha \|\nabla f(x)\| \right] \geq 1 - \delta_S$$

The $\mathbb{P}[\cdot]$ gives the probability of a random variable

Reusing past information to construct subspaces

Suppose D_k has the form $D_k = [D_k^U \ D_k^R] \in \mathbb{R}^{n \times p}$, where

- $D_k^U \in \mathbb{R}^{n \times (p - p_{\text{rand}})}$ is picked from D_{k-1}
- $D_k^R \in \mathbb{R}^{n \times p_{\text{rand}}}$ is randomly generated

Reusing past information to construct subspaces

Suppose D_k has the form $D_k = [D_k^U \ D_k^R] \in \mathbb{R}^{n \times p}$, where

- $D_k^U \in \mathbb{R}^{n \times (p - p_{\text{rand}})}$ is picked from D_{k-1}
- $D_k^R \in \mathbb{R}^{n \times p_{\text{rand}}}$ is randomly generated
- D_k^U is picked such that $\|(D_k^U)^\dagger\|$ is as small as possible
- D_k^R consists of orthogonal columns and $\text{col}(D_k^R) \subseteq \text{col}(D_k^U)^\perp$

Constructing D_k

Recall: $\kappa_f, \kappa_g, \kappa_h$ monotonically increasing w.r.t. $\|D_k^\dagger\|$

Constructing D_k

Recall: $\kappa_f, \kappa_g, \kappa_h$ monotonically increasing w.r.t. $\|D_k^\dagger\|$

Idea: Minimize $\|D_k^\dagger\| = 1/\sigma_{\min}(D_k) \Leftrightarrow$ Maximize $\sigma_{\min}(D_k)$

Constructing D_k

Recall: $\kappa_f, \kappa_g, \kappa_h$ monotonically increasing w.r.t. $\|D_k^\dagger\|$

Idea: Minimize $\|D_k^\dagger\| = 1/\sigma_{\min}(D_k) \Leftrightarrow$ Maximize $\sigma_{\min}(D_k)$

Theorem. Let $\tilde{D} = [d_1 \cdots d_{q-1}] \in \mathbb{R}^{n \times (q-1)}$ where $2 \leq q \leq n$

Define $D(x) : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times q}$ by $D(x) = [\tilde{D} \ x]$

Then for all $x \in \mathbb{R}^n$ with $\|x\| = \Delta$, we have

$$\sigma_{\min}(D(x)) \leq \min \{ \sigma_{\min}(\tilde{D}), \Delta \}.$$

In particular, if $x^* \in \mathbb{R}^n$ with $\|x^*\| = \Delta$ satisfies $d_i^\top x^* = 0$ for all $d_i \in \tilde{D}$, then

$$\sigma_{\min}(D(x^*)) = \min \{ \sigma_{\min}(\tilde{D}), \Delta \}.$$

Constructing D_k^U

Idea: $\sigma_{\min}(D_k^U)$ should be as large as possible

Select p directions to previous sample points as the columns of D_k^U

for removed = 1, ..., p_{rand} **do**

Denote $D_k^U = [d_1^U \cdots d_m^U]$

for $i = 1, \dots, m$ **do**

Define $M_i = [d_1^U \cdots d_{i-1}^U \quad d_{i+1}^U \cdots d_m^U]$ and compute

$$\theta_i = \sigma_{\min}(M_i) \cdot \max\left(\frac{\|d_i^U\|^4}{\Delta_{k+1}^4}, 1\right)$$

Remove the direction with the largest θ_i from D_k^U

Constructing D_k^R

Idea: D_k^R consists of orthogonal columns and $\text{col}(D_k^R) \subseteq \text{col}(D_k^U)^\perp$

Generate $A \in \mathbb{R}^{n \times p_{\text{rand}}}$ with $A_{ij} \sim \mathcal{N}(0, 1/p_{\text{rand}})$

Factorize $D_k^U = QR$ and calculate $\tilde{A} = A - QQ^\top A$

Factorize $\tilde{A} = \tilde{Q}\tilde{R}$

Return $\Delta_k \tilde{Q}$

Theorem. Suppose $f \in \mathcal{C}^1$, $\alpha, \delta_S \in (0, 1)$ and $p_{\text{rand}} \geq 4(1 - \alpha)^{-2} \ln(1/\delta_S)$. Then there exists $\alpha_D > 0$ such that D_k is α_D -well-aligned for f at x_k with probability at least $1 - \delta_S$.

for $k = 0, 1, \dots$ **do**

Define an affine subspace $x_k + D_k \mathbb{R}^P$ by selecting $D_k = [D_k^U \ D_k^R]$

Construct a Q_k -fully quadratic model \widehat{m}_k in \mathbb{R}^P

Apply the remaining part of the random subspace MBTR algorithm

Outline

- 1 Introduction
- 2 Quadratic approximation random subspace trust-region algorithm
- 3 Convergence analysis**
- 4 Numerical experiments
- 5 Summary

Assumptions

- $f \in \mathcal{C}^{2+}$ and bounded below
- $\|\nabla^2 \widehat{m}_k\| \leq \kappa_H$ for all k
- Solution \widehat{s}_k of the trust-region subproblem satisfy

$$\widehat{m}_k(\mathbf{0}) - \widehat{m}_k(\widehat{s}_k) \geq \frac{1}{2} \|\nabla \widehat{m}_k(\mathbf{0})\| \min \left(\Delta_k, \frac{\|\nabla \widehat{m}_k(\mathbf{0})\|}{\max \{\|\nabla^2 \widehat{m}_k\|, 1\}} \right)$$

- Each D_k is α_D -well aligned for f at x_k with probability at least $1 - \delta_S$

Theorem. For all $\epsilon > 0$, there exist $C > 0$ and a sufficiently large K s.t.

$$\mathbb{P} \left[\min_{k \leq K} \|\nabla f(x_k)\| \leq \epsilon \right] \geq 1 - e^{-C(K+1)}$$

The $\mathbb{P}[\cdot]$ and $\mathbb{E}[\cdot]$ give the probability and expected value of a random variable

Theorem. For all $\epsilon > 0$, there exist $C > 0$ and a sufficiently large K s.t.

$$\mathbb{P} \left[\min_{k \leq K} \|\nabla f(x_k)\| \leq \epsilon \right] \geq 1 - e^{-C(K+1)}$$

Theorem. If QARSTA is run with $\Delta_{\min} = 0$, then

$$\mathbb{P} \left[\inf_{k \geq 0} \|\nabla f(x_k)\| = 0 \right] = 1$$

Convergence results

Theorem. For all $\epsilon > 0$, there exist $C > 0$ and a sufficiently large K s.t.

$$\mathbb{P} \left[\min_{k \leq K} \|\nabla f(x_k)\| \leq \epsilon \right] \geq 1 - e^{-C(K+1)}$$

Theorem. If QARSTA is run with $\Delta_{\min} = 0$, then

$$\mathbb{P} \left[\inf_{k \geq 0} \|\nabla f(x_k)\| = 0 \right] = 1$$

Theorem. $\mathbb{E} [\min \{k : \|\nabla f(x_k)\| \leq \epsilon\}] = \mathcal{O}(\epsilon^{-2})$

The $\mathbb{P}[\cdot]$ and $\mathbb{E}[\cdot]$ give the probability and expected value of a random variable

Outline

- 1 Introduction
- 2 Quadratic approximation random subspace trust-region algorithm
- 3 Convergence analysis
- 4 Numerical experiments**
- 5 Summary

- Is it beneficial to use quadratic models instead of linear models?
- What is a good choice of p and p_{rand} ?
- Exploiting the structure of objective functions?

Two sets from the CUTEst collection [Gould, Orban, Toint, 2015] with dimension $n \approx 1000$:

- 73 unconstrained problems with various objective functions
- 32 unconstrained nonlinear least-squares problems, i.e.,

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} \sum_{i=1}^m g_i(x)^2$$

Models:

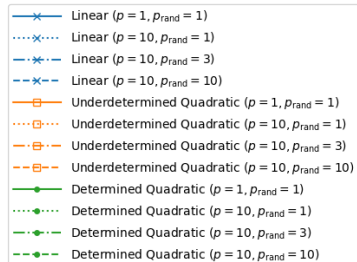
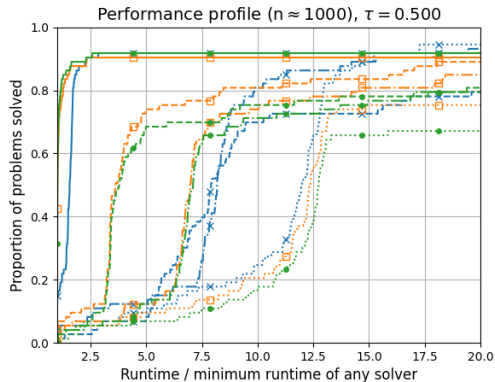
- Linear model ($p + 1$ fevals)
- Underdetermined quadratic model ($2p + 1$ fevals)
- Determined quadratic model ($(p + 1)(p + 2)/2$ fevals)
- Square-of-linear model ($p + 1$ fevals) *On test set 2 only

Parameters:

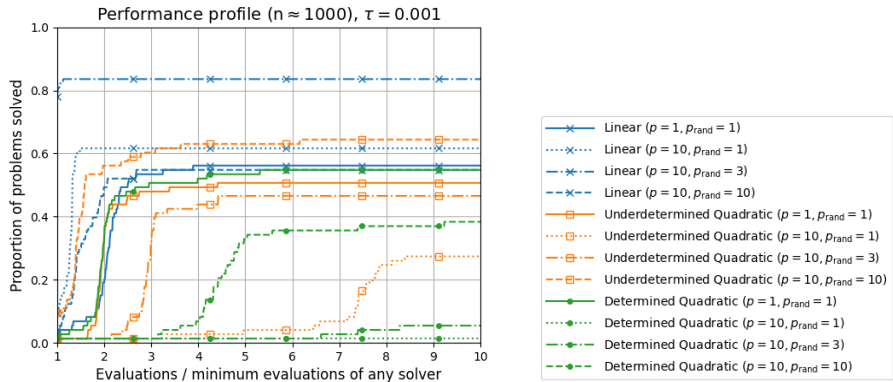
- $p = 1, p_{\text{rand}} = 1$
- $p = 10, p_{\text{rand}} = 1$
- $p = 10, p_{\text{rand}} = 3$
- $p = 10, p_{\text{rand}} = 10$

Success is defined as finding x_k such that $f(x_k) \leq f(x^*) + \tau (f(x_0) - f(x^*))$ within $100(n + 1)$ fevals

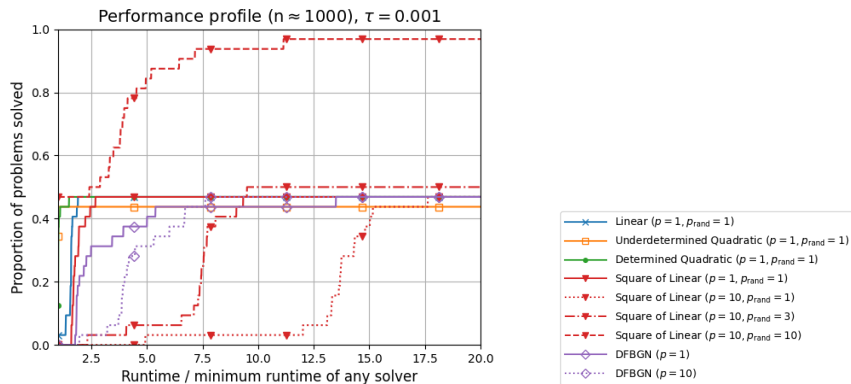
Comparing linear and quadratic models based on runtime



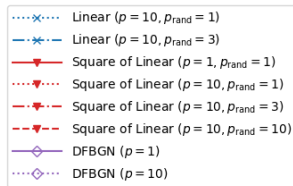
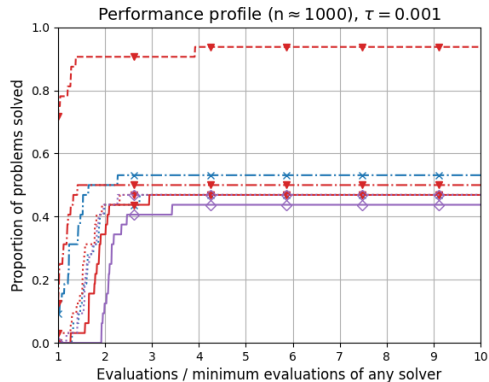
Comparing linear and quadratic models based on evals



Performance on nonlinear least-squares problems (runtime)



Performance on nonlinear least-squares problems (evals)



Outline

- 1 Introduction
- 2 Quadratic approximation random subspace trust-region algorithm
- 3 Convergence analysis
- 4 Numerical experiments
- 5 Summary**

Summary

In this research, we:

- Provided a Q -fully quadratic modeling technique that is easy to analyze and implement
- Proposed an algorithm with convergence analysis for general unconstrained DFO problems
- Demonstrated efficiency of quadratic models and exploiting the structure of objective functions

Future directions:

- Use a hybrid of linear and quadratic models
- Handle constrained DFO problems

Thank you

- <https://github.com/yiwchen233/QARSTA>
- Y. Chen, W. Hare, and A. Wiebe. “Q-fully quadratic modeling and its application in a random subspace derivative-free method”. In: *Computational Optimization and Applications* (2024)