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

Yiwen Chen

Department of Mathematics
University of British Columbia

April, 2024

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

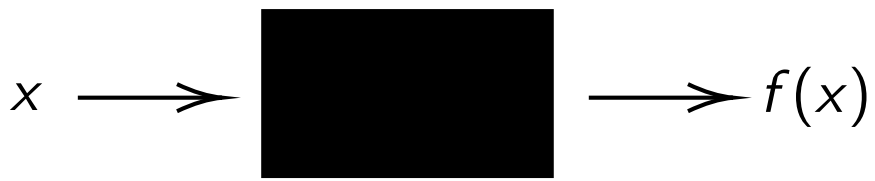
Outline

Optimization with blackbox objective function

Consider the optimization problem

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

where f is given by a blackbox:



Example

Consider a chemical process:

$$\left.\begin{array}{c}\text{Temperature}\\ \text{Humidity}\\ \text{Reaction time}\end{array}\right\} = x \longrightarrow \boxed{} \longrightarrow f(x) = \text{Product purity}$$

Goal:

$$\max\{f(x) : x_1 \in [273, 373], x_2 \in [30, 60], x_3 \in [1, 10]\}$$

Another example

Consider a computer simulation of earthquakes:

$$\left.\begin{array}{c}\text{Height}\\ \text{Material}\\ \text{Structure}\end{array}\right\} = x \longrightarrow \boxed{\phantom{\text{Simulation}}} \longrightarrow f(x) = \text{Stability of building}$$

Goal:

$$\max\{f(x) : x \text{ satisfies some constraints}\}$$

Derivative-free optimization (DFO)

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

Two categories of DFO methods

Direct search methods

- Maintain an incumbent solution and check a finite number of trial points for potential decrease
- E.g., Coordinate Search, MADS

Two categories of DFO methods

Direct search methods

- Maintain an incumbent solution and check a finite number of trial points for potential decrease
- E.g., Coordinate Search, MADS

Model-based methods

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

Polynomial interpolation

Definition. For a given function $f(x)$ and set $Y = \{y^0, ..., y^s\}$, a polynomial $m(x)$ is a polynomial interpolation model of $f(x)$ if

$$m(y^i) = f(y^i), \quad i = 0, ..., s$$

Polynomial interpolation

Definition. For a given function $f(x)$ and set $Y = \{y^0, ..., y^s\}$, a polynomial $m(x)$ is a polynomial interpolation model of $f(x)$ if

$$m(y^i) = f(y^i), \quad i = 0, ..., s$$

Note: In practice, $m(x)$ is determined by finding $\alpha_0, ..., \alpha_t$ such that

$$m(y^i) = \sum_{j=0}^{t} \alpha_j \phi_j(y^i) = f(y^i), \quad i = 0, ..., s$$

where the set of functions $\phi = \{\phi_0, ..., \phi_t\}$ is a polynomial basis

Linear interpolation model

Let $\phi = \{1, x_1, x_2, ..., x_d\}$ and $Y = \{y^0, ..., y^d\}$
Then a linear interpolation model $m(x)$ is determined by finding $\alpha_0, ..., \alpha_d$ such that

$$m(y^i) = \alpha_0 + \alpha_1 y_1^i + \cdots + \alpha_d y_d^i = f(y^i), \quad i = 0, ..., d$$

where y_j^i is the j-th element of y^i

Quadratic interpolation model

Let $\phi = \{1, x_1, x_2, ..., x_d, \frac{x_1^2}{2}, x_1x_2, ..., x_{d-1}x_d, \frac{x_d^2}{2}\}$ and $Y = \{y^0, ..., y^s\}$
Then a quadratic interpolation model $m(x)$ is determined by finding $\alpha_0, ..., \alpha_{\frac{(d+1)(d+2)}{2}-1}$ such that

$$m(y^i) = \alpha_0 + \alpha_1 y_1^i + \cdots + \alpha_d y_d^i + \alpha_{d+1}\frac{(y_1^i)^2}{2} + \cdots + \alpha_{\frac{(d+1)(d+2)}{2}-1}\frac{(y_d^i)^2}{2} = f(y^i), \quad i = 0, ..., s$$

Quadratic interpolation model

Let $\phi = \{1, x_1, x_2, ..., x_d, \frac{x_1^2}{2}, x_1x_2, ..., x_{d-1}x_d, \frac{x_d^2}{2}\}$ and $Y = \{y^0, ..., y^s\}$
Then a quadratic interpolation model $m(x)$ is determined by finding $\alpha_0, ..., \alpha_{\frac{(d+1)(d+2)}{2}-1}$ such that

$$m(y^i) = \alpha_0 + \alpha_1 y_1^i + \cdots + \alpha_d y_d^i + \alpha_{d+1}\frac{(y_1^i)^2}{2} + \cdots + \alpha_{\frac{(d+1)(d+2)}{2}-1}\frac{(y_d^i)^2}{2} = f(y^i), \quad i = 0, ..., s$$

Note: If $s = \frac{(d+1)(d+2)}{2} - 1$ and the system has full rank, then $m(x)$ is called a determined quadratic interpolation model

If $s < \frac{(d+1)(d+2)}{2} - 1$ and the system has full rank, then $m(x)$ is called an underdetermined quadratic interpolation model

Model-based DFO: limitations

Limitations:

- Function evaluations are too expensive for large problems

d	1	10	100	1000
$(d+1)(d+2)/2$	3	66	5151	501501

- 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]

The inspirational question

Is it beneficial to use quadratic models instead of linear models?

Outline

Model-based trust-region (MBTR) algorithm

for $k = 0, 1, \ldots$ **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 $\mathbb{R}^n$:

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

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

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

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

Random subspace MBTR algorithm

for $k = 0, 1, \ldots$ **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 $\mathbb{R}^p$:

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

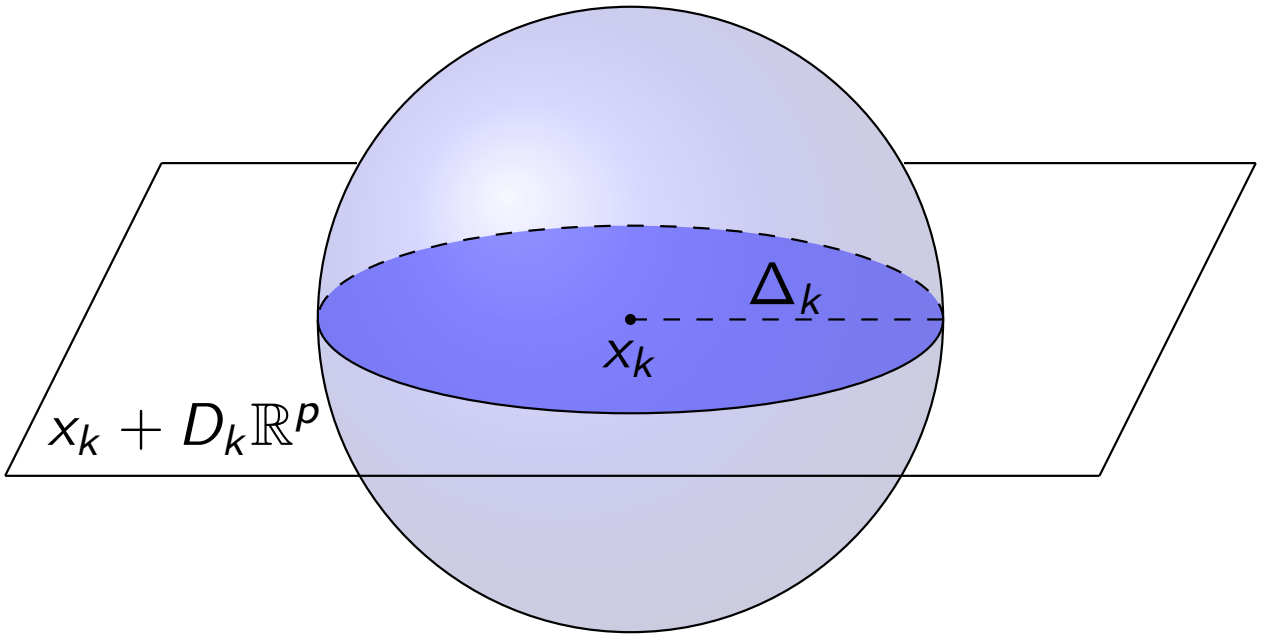
and calculate the corresponding step $s_k \in \mathbb{R}^n$

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

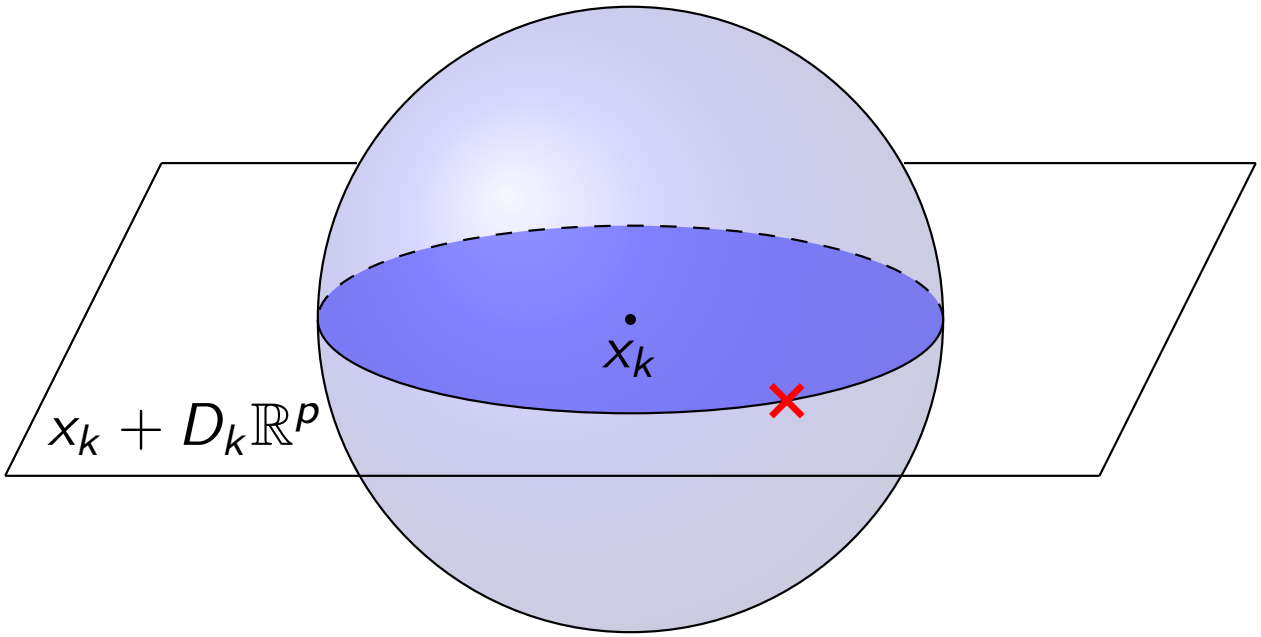
$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{\widehat{m}_k(\mathbf{0}) - \widehat{m}_k(\widehat{s}_k)} = \frac{\textit{true decrease}}{\textit{predicted decrease}}$$

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

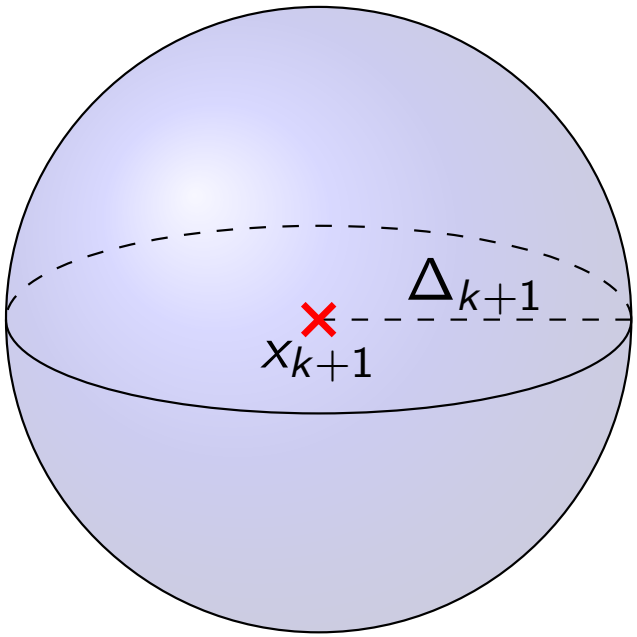
Random subspace MBTR algorithm: visuals



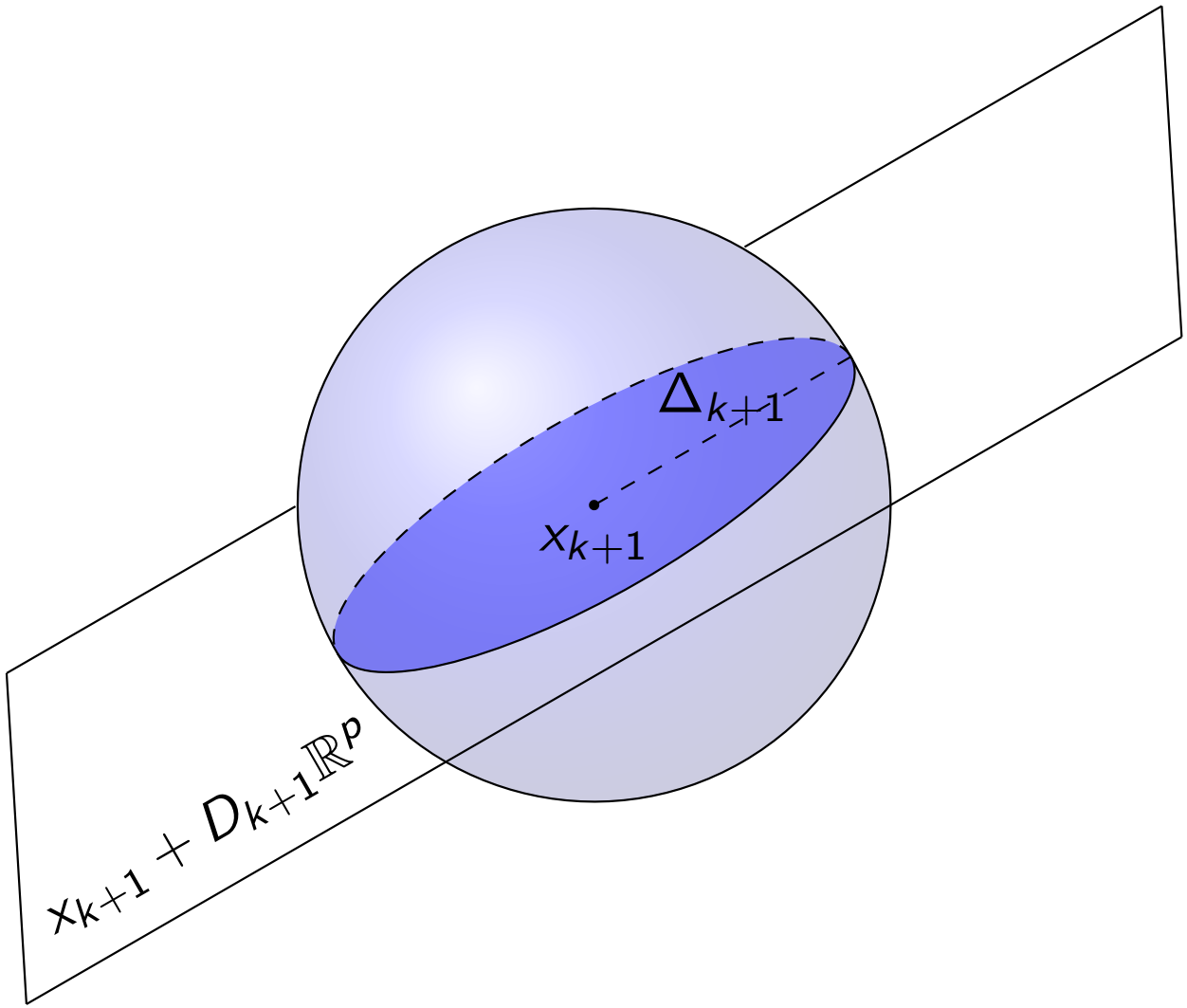
Random subspace MBTR algorithm: visuals



Random subspace MBTR algorithm: visuals



Random subspace MBTR algorithm: visuals



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

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

Generalized simplex gradient

Definition. [Hare, Jarry-Bolduc, 2020]
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

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

where

$$\delta_f(x^0; D) = \begin{bmatrix} f(x^0 + d_1) - f(x^0) \\ f(x^0 + d_2) - f(x^0) \\ \vdots \\ f(x^0 + d_p) - f(x^0) \end{bmatrix}$$

Generalized simplex Hessian

Definition. [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 Hessian of f at x^0 over D is defined by

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

where

$$\delta_{\nabla_S f}(x^0; D) = \begin{bmatrix} \left(\nabla_S f(x^0 + d_1; D) - \nabla_S f(x^0; D)\right)^\top \\ \left(\nabla_S f(x^0 + d_2; D) - \nabla_S f(x^0; D)\right)^\top \\ \vdots \\ \left(\nabla_S f(x^0 + d_p; D) - \nabla_S f(x^0; D)\right)^\top \end{bmatrix}$$

Constructing Q-fully quadratic models

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

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

The model $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ is defined by

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

The $2\nabla_S f(x^0; D) - \nabla_S f(x^0; 2D)$ is a special case of the *Adapted Centred Simplex Gradient*, see Y. Chen and W. Hare. "Adapting the centred simplex gradient to compensate for misaligned sample points". In: *IMA J. Numer. Anal.* (2023)

Constructing Q-fully quadratic models

Theorem. $\widehat{m}$ is a determined quadratic interpolation model of $f(x^0 + Q\widehat{s})$

Constructing Q-fully quadratic models

Theorem. $\widehat{m}$ is a determined quadratic interpolation model of $f(x^0 + Q\widehat{s})$

Theorem. 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 $\kappa_{ef}, \kappa_{eg}, \kappa_{eh}$ monotonically increasing w.r.t. $\|D^\dagger\|$

Subspace selection

α-well-aligned matrices

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

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

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

Can we reuse 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

Can we reuse 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 $\sigma_{\min}(D_k^U)$ is as large 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: $\kappa_{ef}, \kappa_{eg}, \kappa_{eh}$ monotonically increasing w.r.t. $\|D_k^\dagger\|$

Constructing D_k

Recall: $\kappa_{ef}, \kappa_{eg}, \kappa_{eh}$ 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: $\kappa_{ef}, \kappa_{eg}, \kappa_{eh}$ 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 $\widetilde{D} = [d_1 \cdots d_{q-1}] \in \mathbb{R}^{n\times(q-1)}$ where $2 \le q \le n$
Define $D(x) : \mathbb{R}^n \to \mathbb{R}^{n\times q}$ by $D(x) = [\widetilde{D}\ x]$
Then for all $x \in \mathbb{R}^n$ with $\|x\| = \Delta$, we have

$$\sigma_{\min}(D(x)) \le \min\left\{\sigma_{\min}(\widetilde{D}), \Delta\right\}.$$

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

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

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 $\texttt{removed} = 1, ..., p_{\text{rand}}$ **do**

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

  for $i = 1, ..., m$ **do**

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

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

  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 $\mathrm{col}(D_k^R) \subseteq \mathrm{col}(D_k^U)^\perp$

Generate $A \in \mathbb{R}^{n \times p_{\mathrm{rand}}}$ with $A_{ij} \sim \mathcal{N}(0, 1/p_{\mathrm{rand}})$
Factorize $D_k^U = QR$ and calculate $\widetilde{A} = A - QQ^\top A$
Factorize $\widetilde{A} = \widetilde{Q}\widetilde{R}$
Return $\Delta_k \widetilde{Q}$

Algorithm modified from C. Cartis and L. Roberts. "Scalable subspace methods for derivative-free nonlinear least-squares optimization". In: *Math. Program.* 199.1-2 (2023), pp. 461–524

Subspace quality

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 $1 - \delta_S$

QARSTA

for $k = 0, 1, \ldots$ **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 $\mathbb{R}^p$

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

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

and calculate the corresponding step $s_k \in \mathbb{R}^n$
Evaluate $f(x_k + s_k)$ and calculate ratio

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

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

Outline

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 $1 - \delta_S$

Convergence results

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

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

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

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

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

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

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

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

Outline

Questions

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

Test problems

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.,

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

Solvers

Models:

- Linear model
- Underdetermined quadratic model
- Determined quadratic model
- Square-of-linear model (on the second problem set only)

Parameters:

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

Performance profiles

For each solver $S \in \mathcal{S}$ and problem $P \in \mathcal{P}$, define the performance ratio

$$r_{P,S} = \frac{t_{P,S}}{\min\{t_{P,S} : S \in \mathcal{S}\}},$$

where $t_{P,S} > 0$ is the performance measure

The performance profile of S is

$$\rho_S(\alpha) = \frac{1}{|\mathcal{P}|} |\{P \in \mathcal{P} : r_{P,S} \leq \alpha\}|$$

Stopping criteria and performance measure

For each solver S and problem P with dimension n_P
Stopping criteria:

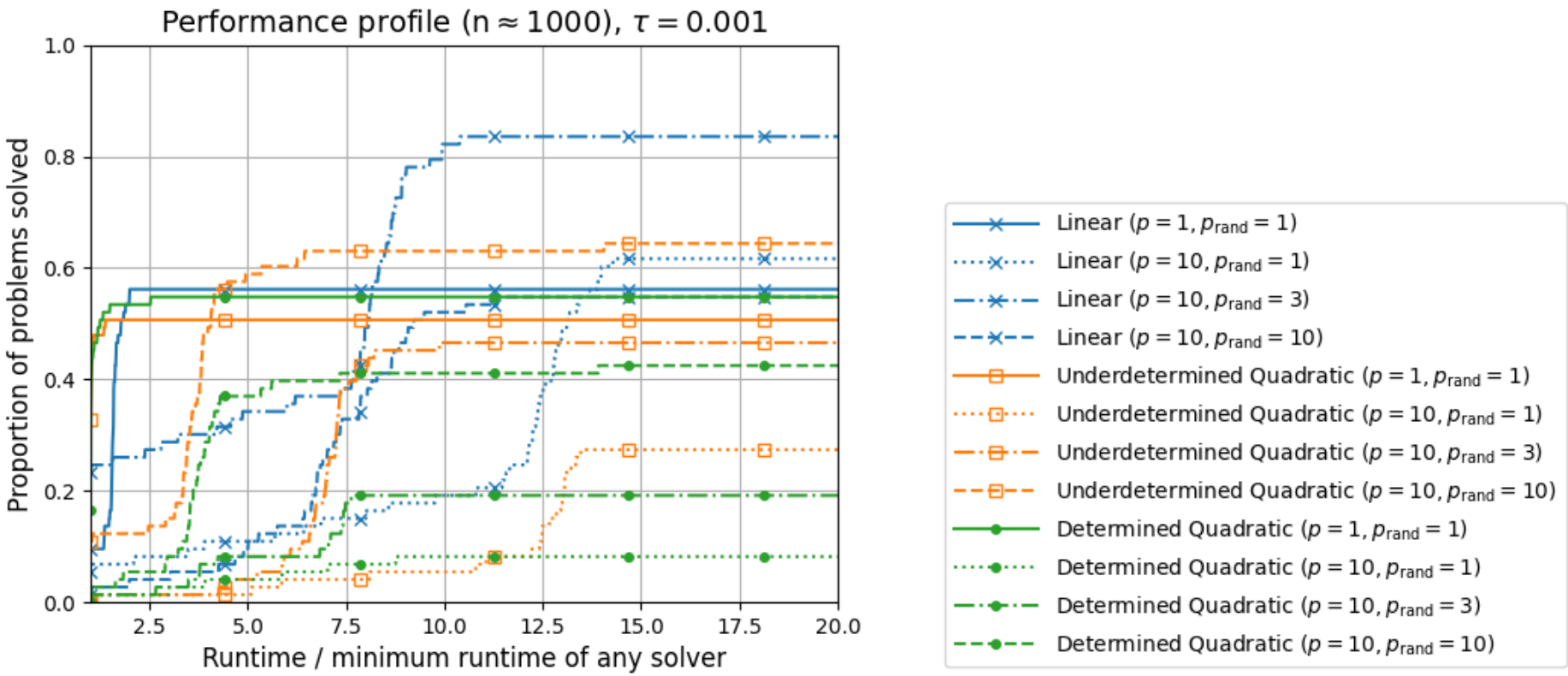
- $f(x_k) \leq f(x^*) + \tau\,(f(x_0) - f(x^*))$ (success)

or

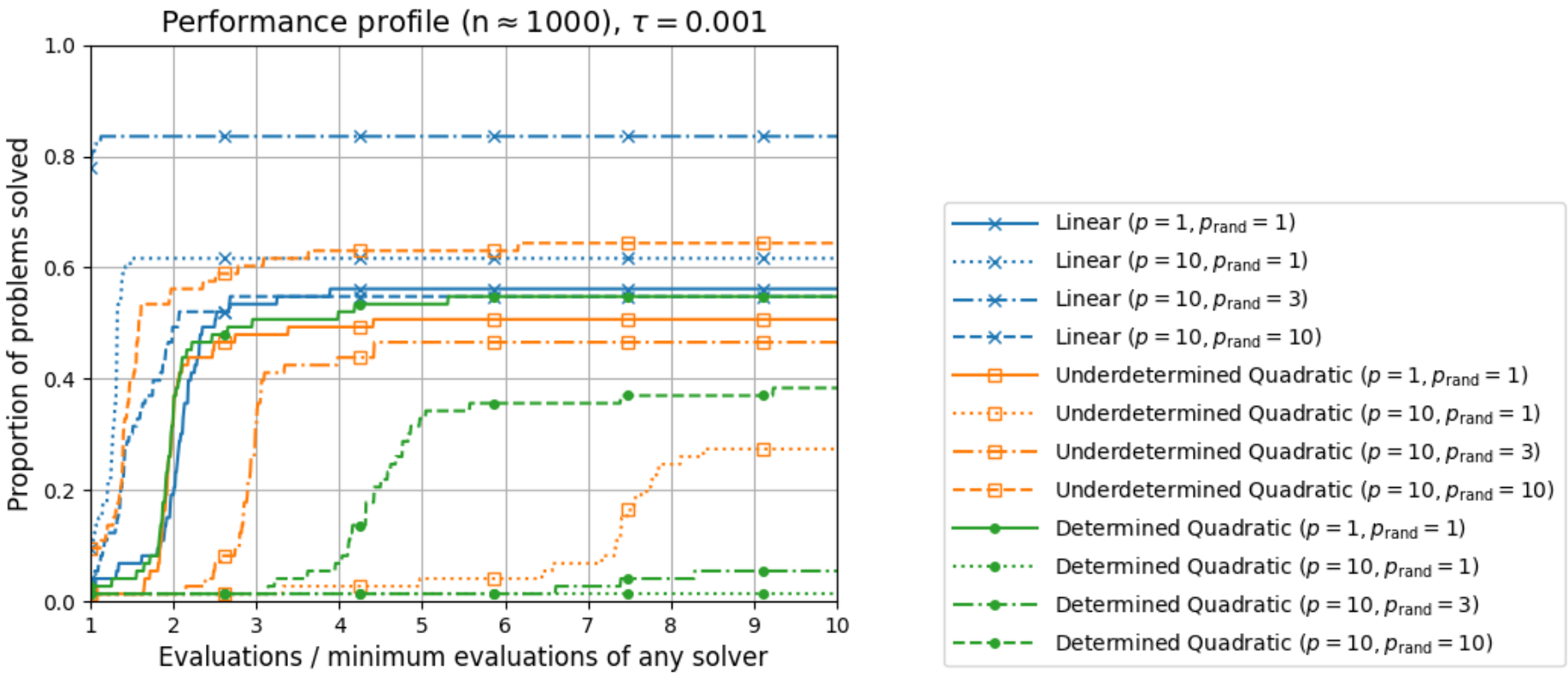
- more than $100(n_P + 1)$ function evaluations are needed (failure)

If success, then $t_{P,S}$ is the number of function evaluations or the runtime
If failure, then $t_{P,S} = \infty$

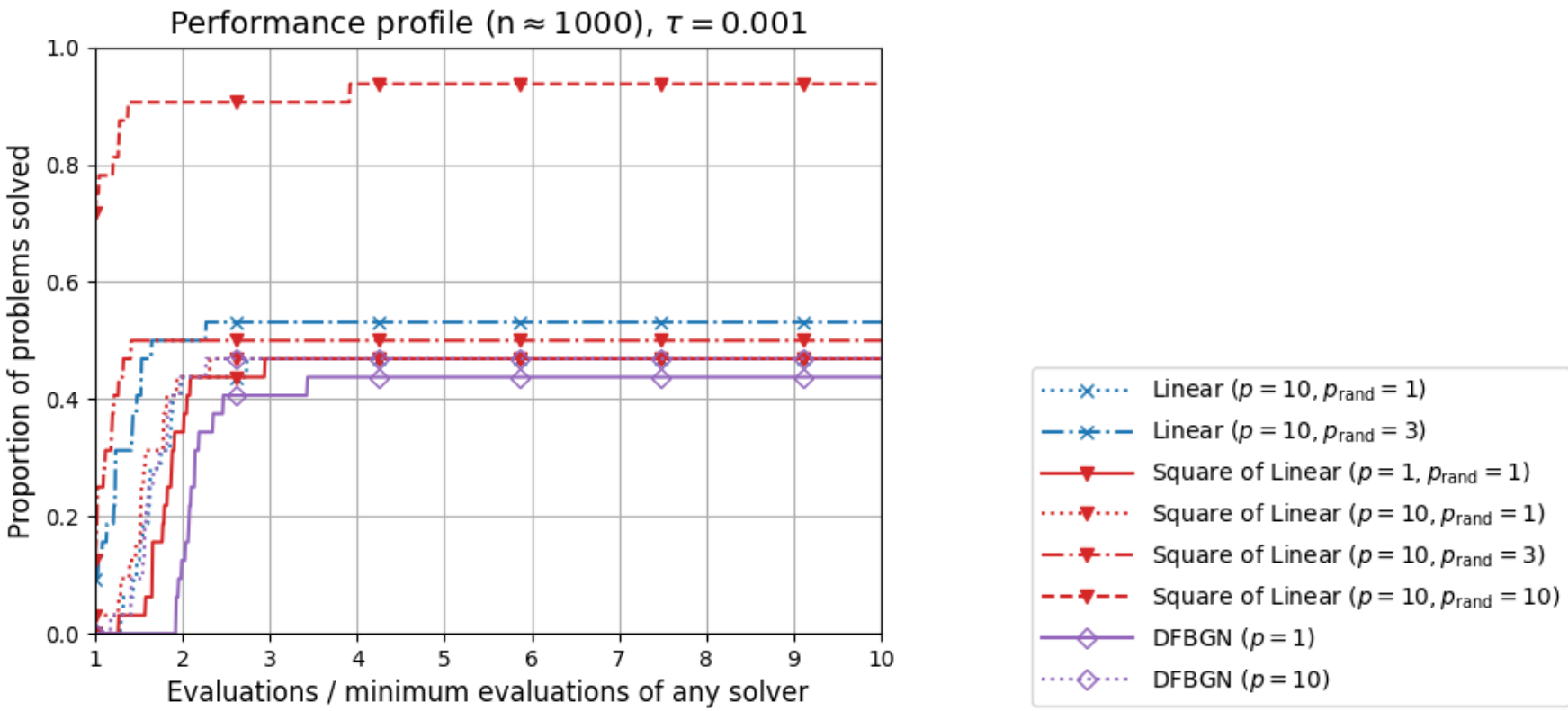
Comparing linear and quadratic models based on runtime



Comparing linear and quadratic models based on evals



Advantages of exploiting the structure of obj functions



Outline

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 the efficiency of using quadratic models and exploiting the structure of objective functions

Future directions:

- Compare with other underdetermined quadratic models
- Design better strategies of selecting p and p_{rand}
- Handle constrained DFO problems

Thank you

- Yiwen Chen, Warren Hare, and Amy Wiebe. "Q-fully Quadratic Modeling and its Application in a Random Subspace Derivative-free Method". In: *arXiv preprint* (2023). URL: https://arxiv.org/abs/2312.03169