Randomized subspace methods for high-dimensional model-based derivative-free optimization

Yiwen Chen

Department of Mathematics
University of British Columbia

December 5, 2025

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

Outline

Consider the optimization problem

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

where f is given by a blackbox:

Example

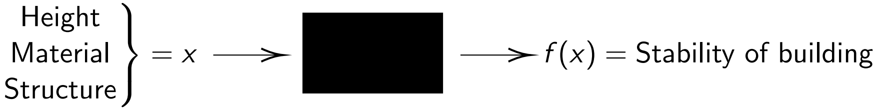
Consider a chemical process:

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:



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, Mesh Adaptive Direct Search

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 $Y = \{y^0, \ldots, y^s\}$, a polynomial $m(x)$ is a polynomial interpolation model of $f(x)$ if

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

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

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

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

Linear interpolation model

Let $\phi = \{1, x_1, x_2, \ldots, x_n\}$ and $Y = \{y^0, \ldots, y^s\} \subseteq \mathbb{R}^n$
Then a linear interpolation model $m(x)$ is determined by finding $\alpha_0, \ldots, \alpha_n$ such that

$$m(y^i) = \alpha_0 + \alpha_1 y_1^i + \cdots + \alpha_n y_n^i = \begin{bmatrix} 1 & (y^i)^\top \end{bmatrix} \alpha = f(y^i), \quad i = 0, \ldots, s$$

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

Note: If $s = n$ and the system has full rank, then $m(x)$ is called a determined linear interpolation model

If $s < n$ and the system has full rank, then $m(x)$ is called an underdetermined linear interpolation model

Quadratic interpolation model

Let $\phi = \{1, x_1, x_2, \ldots, x_n, \frac{x_1^2}{2}, x_1 x_2, \ldots, x_{n-1}x_n, \frac{x_n^2}{2}\}$ and
$Y = \{y^0, \ldots, y^s\} \subseteq \mathbb{R}^n$
Then a quadratic interpolation model $m(x)$ is determined by finding
$\alpha_0, \ldots, \alpha_{\frac{(n+1)(n+2)}{2}-1}$ such that

$$m(y^i) = \alpha_0 + \alpha_1 y_1^i + \cdots + \alpha_n y_n^i + \alpha_{n+1}\frac{\left(y_1^i\right)^2}{2} + \cdots + \alpha_{\frac{(n+1)(n+2)}{2}-1}\frac{\left(y_n^i\right)^2}{2} = f(y^i), \quad i = 0, \ldots, s$$

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

Model-based DFO: limitations

Limitations:

- Number of function evals. is too high for large problems ($n \approx 1000$)

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

- Primarily designed for small- to medium-scale problems ($n \leq 100$)

Model-based DFO through subspaces

Idea:

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

Some existing papers:
[Zhang, 2012]; [Cartis, Roberts, 2023]; [Dzahini, Wild, 2024];
[Chen, Hare, Wiebe, 2024]; [Cartis, Roberts, 2024]
[Chen, Hare, Wiebe, 2025]

Outline

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 $\mathbb{R}^n$:

$$s^k \approx \operatorname*{argmin}_{s \in \mathbb{R}^n} m^k(s), \quad s.t. \ \|s\| \leq \Delta^k$$

Evaluate $f(x^k + s^k)$ and apply descent ratio test

$$\rho^k = \frac{f(x^k) - f(x^k + s^k)}{m^k(\mathbb{0}) - m^k(s^k)} = \frac{true \ decrease}{predicted \ decrease}$$

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 $\mathbb{R}^p$:

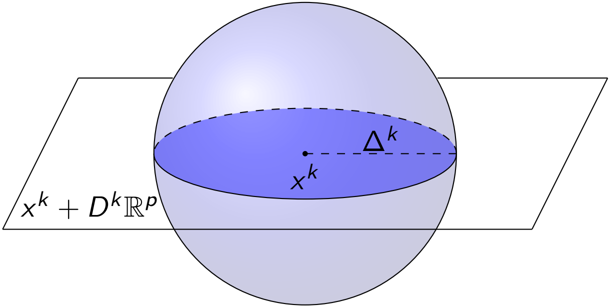
$$\widehat{s}^k \approx \operatorname*{argmin}_{\widehat{s} \in \mathbb{R}^p} \widehat{m}^k(\widehat{s}), \quad s.t. \; \|\widehat{s}\| \leq \Delta^k$$

and calculate the corresponding step $s^k \in \mathbb{R}^n$
Evaluate $f(x^k + s^k)$ and apply descent ratio test

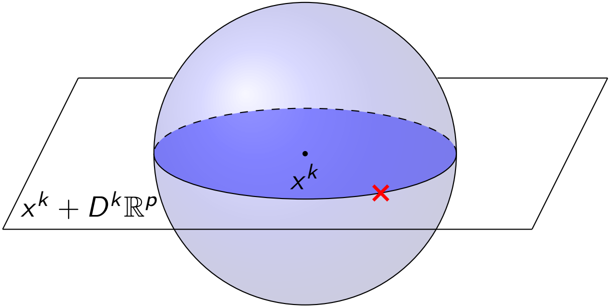
$$\rho^k = \frac{f(x^k) - f(x^k + s^k)}{\widehat{m}^k(\mathbb{0}) - \widehat{m}^k(\widehat{s}^k)} = \frac{\textit{true decrease}}{\textit{predicted decrease}}$$

Accept/reject step based on ρ^k and update trust region radius

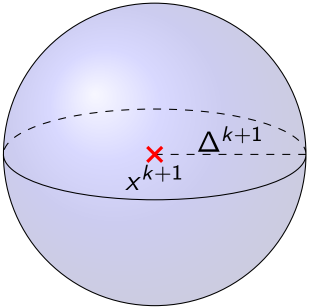
Random subspace MBTR algorithm: visuals



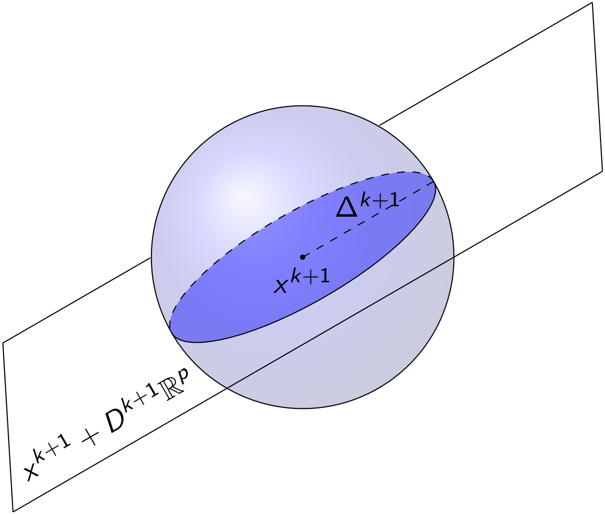
Random subspace MBTR algorithm: visuals



Random subspace MBTR algorithm: visuals



Random subspace MBTR algorithm: visuals



Model construction

Q-fully linear models

Definition. A model $m : \mathbb{R}^n \to \mathbb{R}$ is fully linear in $B(x, \Delta)$ if there exist $\kappa_f, \kappa_g > 0$ s.t. for all $s \in \mathbb{R}^n$ with $\|s\| \leq \Delta$,

$$|f(x+s) - m(x+s)| \leq \kappa_f \Delta^2$$
$$\|\nabla f(x+s) - \nabla m(x+s)\| \leq \kappa_g \Delta$$

Q-fully linear models

Definition. A model $m : \mathbb{R}^n \to \mathbb{R}$ is fully linear in $B(x, \Delta)$ if there exist $\kappa_f, \kappa_g > 0$ s.t. for all $s \in \mathbb{R}^n$ with $\|s\| \leq \Delta$,

$$
\begin{aligned}
|f(x+s) - m(x+s)| &\leq \kappa_f \Delta^2 \\
\|\nabla f(x+s) - \nabla m(x+s)\| &\leq \kappa_g \Delta
\end{aligned}
$$

Definition. [Cartis, Roberts, 2023]
Let $Q \in \mathbb{R}^{n \times p}$. A model $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ is Q-fully linear in $B(x, \Delta)$ if there exist $\kappa_f, \kappa_g > 0$ s.t. for all $\widehat{s} \in \mathbb{R}^p$ with $\|\widehat{s}\| \leq \Delta$,

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

Q-fully quadratic models

Definition. [Chen, Hare, Wiebe, 2024]
Let $Q \in \mathbb{R}^{n\times p}$. A model $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ is Q-fully quadratic in $B(x,\Delta)$ if there exist $\kappa_f, \kappa_g, \kappa_h > 0$ s.t. for all $\widehat{s} \in \mathbb{R}^p$ with $\|\widehat{s}\| \leq \Delta$,

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

Generalized simplex gradient

Definition. [Custódio, Dennis Jr., Vicente, 2008]
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

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

where

$$\delta_{\nabla_s f}(x^0; D; D) = \begin{bmatrix} \left(\nabla_s f(x^0 + d^1; D) - \nabla_s f(x^0; D)\right)^\top \\ \left(\nabla_s f(x^0 + d^2; D) - \nabla_s f(x^0; D)\right)^\top \\ \vdots \\ \left(\nabla_s f(x^0 + d^p; D) - \nabla_s f(x^0; D)\right)^\top \end{bmatrix}$$

Constructing Q-fully linear models

Theorem. [Chen, Hare, Wiebe, 2024]
Let $x^0 \in \mathbb{R}^n$ and $D = QR$ have full col. rank, where $R = [r^1 \cdots r^p] \in \mathbb{R}^{p \times p}$
Let

$$m(x) = f(x^0) + \nabla_s f(x^0; D)^\top \left(x - x^0\right)$$

Then, the model $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ defined by $\widehat{m}(\widehat{s}) = m(x^0 + Q\widehat{s})$ is Q-fully linear in $B(x^0, \max_{1 \leq i \leq p} \|r^i\|)$

Moreover, κ_f and κ_g are monotonically increasing w.r.t. $\|D^\dagger\|$

Constructing Q-fully quadratic models

Theorem. [Chen, Hare, Wiebe, 2024]
Let $x^0 \in \mathbb{R}^n$ and $D = QR$ have full col. rank, where $R = [r^1 \cdots r^p] \in \mathbb{R}^{p \times p}$
Let

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

Then, the model $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ defined by $\widehat{m}(\widehat{s}) = m(x^0 + Q\widehat{s})$ is Q-fully quadratic in $B(x^0, \max_{1 \le i \le p} \|r^i\|)$
Moreover, $\kappa_f, \kappa_g, \kappa_h$ are monotonically increasing w.r.t. $\|D^\dagger\|$

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)

Subspace selection

α-well-aligned matrices

Let $x + D\mathbb{R}^p$ be the affine subspace

Definition. [Cartis, Roberts, 2023]
Let $\alpha \in (0,1)$. We say that $D \in \mathbb{R}^{n \times p}$ is α-well-aligned for f at x if

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

α-well-aligned matrices

Let $x + D\mathbb{R}^p$ be the affine subspace

Definition. [Cartis, Roberts, 2023]
Let $\alpha \in (0,1)$. We say that $D \in \mathbb{R}^{n \times p}$ is α-well-aligned for f at x if

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

Theorem. [Dzahini, Wild, 2024] (Idea: Johnson–Lindenstrauss Lemma)
Let $\alpha, \delta \in (0,1)$. Suppose $p \geq 4(1-\alpha)^{-2} \ln(1/\delta)$ and let $D_{ij} \sim \mathcal{N}(0, 1/p)$
Then,

$$\mathbb{P}\left[D \text{ is } \alpha\text{-well-aligned for } f \text{ at x}\right] \geq 1 - \delta$$

Can we reuse past information to construct subspaces?

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

- $D^U \in \mathbb{R}^{n \times (p - p_{\mathrm{rand}})}$ is picked from previous sample points
- $D^R \in \mathbb{R}^{n \times p_{\mathrm{rand}}}$ is randomly generated

Can we reuse past information to construct subspaces?

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

- $D^U \in \mathbb{R}^{n \times (p - p_{\mathrm{rand}})}$ is picked from previous sample points
- $D^R \in \mathbb{R}^{n \times p_{\mathrm{rand}}}$ is randomly generated
- D^U is picked such that $\sigma_{\min}(D^U)$ is as large as possible
- D^R consists of orthogonal columns and $\mathrm{col}(D^R) \subseteq \mathrm{col}(D^U)^{\perp}$

Constructing D

Recall: $\kappa_f, \kappa_g, \kappa_h$ monotonically increasing w.r.t. $\|D^\dagger\|$
Idea: Minimize $\|D^\dagger\| = 1/\sigma_{\min}(D) \Leftrightarrow$ Maximize $\sigma_{\min}(D)$

Constructing D

Recall: $\kappa_f, \kappa_g, \kappa_h$ monotonically increasing w.r.t. $\|D^\dagger\|$
Idea: Minimize $\|D^\dagger\| = 1/\sigma_{\min}(D) \Leftrightarrow$ Maximize $\sigma_{\min}(D)$

Theorem. [Chen, Hare, Wiebe, 2024]
Let $\widetilde{D} = [d^1 \cdots d^{q-1}] \in \mathbb{R}^{n\times(q-1)}$ where $2 \leq q \leq n$
Define $D(x) : \mathbb{R}^n \to \mathbb{R}^{n\times q}$ by $D(x) = [\widetilde{D}\ x]$
Then, for all $x \in \mathbb{R}^n$ with $\|x\| = \Delta$,

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

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

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

Subspace quality

Theorem. [Chen, Hare, Wiebe, 2024]
Let $\alpha, \delta \in (0, 1)$
Suppose $p_{\text{rand}} \geq 4(1-\alpha)^{-2} \ln(1/\delta)$ and let

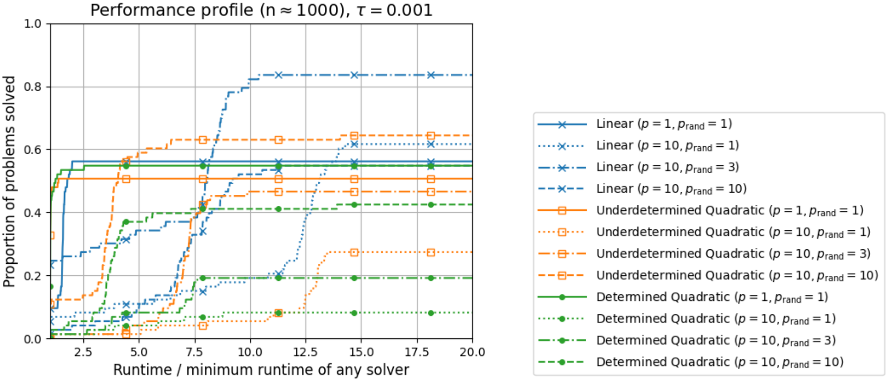
$$D^R \sim \operatorname{proj}_{\operatorname{col}(D^U)^\perp}\left(\mathcal{MN}\left(\mathbb{0}, \frac{1}{p} I_n, I_{p_{\text{rand}}}\right)\right)$$

Then, there exists $\alpha_D > 0$ such that

$$\mathbb{P}\left[D \text{ is } \alpha_D\text{-well-aligned for } f \text{ at x}\right] \geq 1 - \delta$$

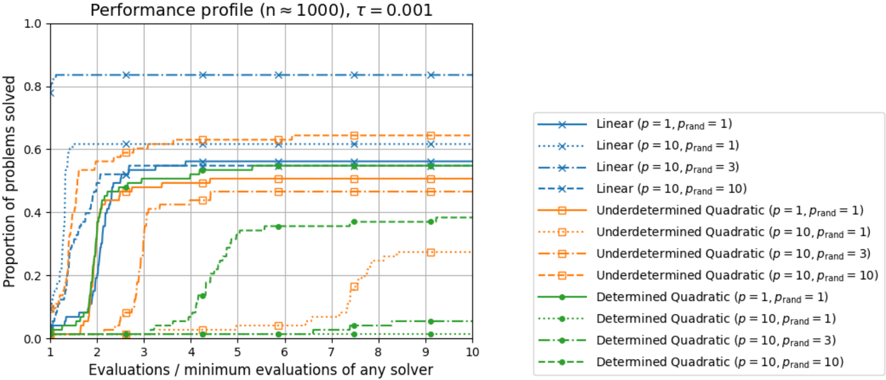
Numerical results

Comparing linear and quadratic models based on runtime



Success is defined as finding x^k such that
$f(x^k) \leq f(x^*) + \tau(f(x^0) - f(x^*))$ in less than $100(n+1)$ f-evals

Comparing linear and quadratic models based on f-evals



Success is defined as finding x^k such that
$f(x^k) \leq f(x^*) + \tau(f(x^0) - f(x^*))$ in less than $100(n+1)$ f-evals

Outline

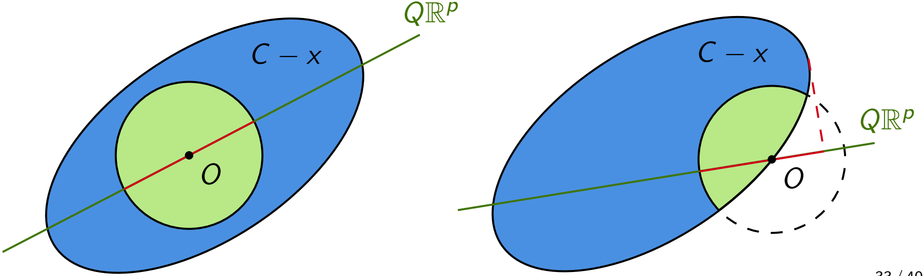
(C, Q)-fully linear models

Let C be the constraint set (convex, closed, nonempty interior)

Definition. [Chen, Hare, Wiebe, 2025]
Let $Q \in \mathbb{R}^{n \times p}$. A model $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ is (C, Q)-fully linear in $B(x, \Delta)$ if there exist $\kappa_f, \kappa_g > 0$ s.t. for all $\widehat{s} \in Q^\top (C - x)$ with $\|\widehat{s}\| \le \Delta$,

$$|f(x + Q\widehat{s}) - \widehat{m}(\widehat{s})| \le \kappa_f \Delta^2$$

$$\max_{\substack{d \in Q^\top (C - x) \\ \|d\| \le 1}} \left| \left(Q^\top \nabla f(x + Q\widehat{s}) - \nabla \widehat{m}(\widehat{s}) \right)^\top d \right| \le \kappa_g \Delta$$



Constructing (C, Q)-fully linear models

Theorem. [Chen, Hare, Wiebe, 2025]
Let $x^0 \in \mathbb{R}^n$ and $D = QR$ have full col. rank, where $R = [r^1 \cdots r^p] \in \mathbb{R}^{p \times p}$
Let

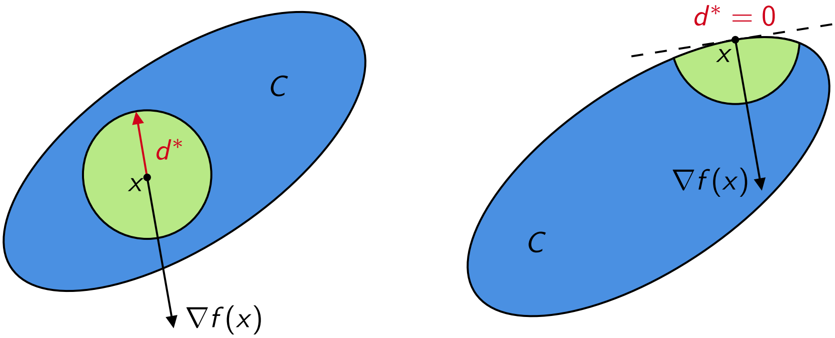
$$m(x) = f(x^0) + \nabla_s f(x^0; D)^\top \left(x - x^0\right)$$

Then, the model $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ defined by $\widehat{m}(\widehat{s}) = m(x^0 + Q\widehat{s})$ is (C, Q)-fully linear in $B(x^0, \max_{1 \le i \le p} \|r^i\|)$

First-order criticality measure

First-order criticality measure for convex-constrained optimization
[Conn, Gould, Toint, 2000]

$$\pi^f(x) = \left| \min_{\substack{x+d \in C \\ \|d\| \leq 1}} \nabla f(x)^\top d \right|$$



α-well-aligned matrices (convex-constrained version)

Let $x + D\mathbb{R}^p$ be the affine subspace and $D = QR$ be the QR factorization

Definition. [Chen, Hare, Wiebe, 2025]
Let $\alpha \in (0,1)$. We say that $D \in \mathbb{R}^{n\times p}$ is α-well-aligned for f and C at x if

$$\left| \min_{\substack{d \in C - x \\ \|d\| \leq 1}} \nabla f(x)^\top Q Q^\top d \right| \geq \alpha \pi^f(x)$$

α-well-aligned matrices (convex-constrained version)

Let $x + D\mathbb{R}^p$ be the affine subspace and $D = QR$ be the QR factorization

Definition. [Chen, Hare, Wiebe, 2025]
Let $\alpha \in (0, 1)$. We say that $D \in \mathbb{R}^{n \times p}$ is α-well-aligned for f and C at x if

$$\left| \min_{\substack{d \in C - x \\ \|d\| \leq 1}} \nabla f(x)^\top Q Q^\top d \right| \geq \alpha \pi^f(x)$$

Theorem. [Chen, Hare, Wiebe, 2025]
(Idea: Concentration on the Grassmannian)
Suppose $p \geq n\alpha$ and let $D_{ij} \sim \mathcal{N}(0, 1)$. Then,

$\mathbb{P}\left[D \text{ is } \alpha\text{-well-aligned for } f \text{ and } C \text{ at x}\right]$
$\geq$ complicated stuff that depends on n, p, α, $\pi^f(x)$, and $\|\nabla f(x)\|$

Convergence and complexity results

Let $\epsilon > 0$, (UC)=UnConstrained, and (CC)=Convex-Constrained

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

  (CC) $\mathbb{P}\left[\min_{k\leq K} \pi^f(x^k) < \epsilon\right] \geq 1 - e^{-C(\epsilon)(K+1)}$

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

  (CC) $\mathbb{P}\left[\inf_{k\geq 0} \pi^f(x^k) = 0\right] = 1$

- (UC) $\mathbb{E}\left[\min\left\{k \geq 0 : \left\|\nabla f(x^k)\right\| < \epsilon\right\}\right] = \mathcal{O}(\epsilon^{-2})$

  (CC) $\mathbb{E}\left[\min\left\{k \geq 0 : \pi^f(x^k) < \epsilon\right\}\right] = \mathcal{O}(\epsilon^{-4})$

(UC) from [Cartis, Roberts, 2023]; (CC) from [Chen, Hare, Wiebe, 2025]

Outline

Summary

In summary,

- High-dimensional DFO problems are hard
- Unconstrained problems can be effectively approached by randomized subspace methods
- Randomized subspace methods work for convex constrained problems, but the projection onto the feasible set is required

Future directions:

- Nonconvex constraints?
- Blackbox constraints?
- Random manifolds?

Thank you

- 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* 89.2 (2024), pp. 317–360
- Y. Chen, W. Hare, and A. Wiebe. "CLARSTA: A random subspace trust-region algorithm for convex-constrained derivative-free optimization". In: *arXiv preprint arXiv:2506.20335* (2025)

Email: `yiwchen@student.ubc.ca`